Metagenomic next generation sequencing for viral pathogens: Application and validation of a deployable sequencer for virus identification

Liana Eleni Kafetzopoulou

July, 2020

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

Authors Declaration

I declare that the work in this thesis was carried out in accordance with the regulations of the University of Liverpool. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree.

Any views expressed in the thesis are those of the author and in no way represent the University of Liverpool.

The thesis has not been presented to any other University for examination either in the United Kingdom or overseas.

Signed:  Date: 07/07/2020

Acknowledgments

It's more than once that people told me a PhD is a very lonely process, having now completed mine I have to admit the last four years were anything but lonely. I gained new colleagues, came closer with older ones and most importantly, after these 4 years found myself with more friends than I ever thought possible. The moment has come to write the last thing left on the thesis to-do list and it might go without saying but none of the things I have included are obvious, expected or assumed; I am a better version of myself and complete this project having gained so much not only in knowledge but also in experiences, people and friends.

First and foremost, I would like to thank Steve Pullan for being such an awesome supervisor, for all the support through the highs and lows and for patiently (emphasis on patiently) guiding me throughout this whole process. I could not have asked for a better supervisor and it has really been an amazing 4 years. I have learned so much working under your guidance and feel really lucky I had such a great time during my PhD. I always felt I could reach out and you did a great job keeping me in check during the whole process, for which I am very grateful and always will be. Big thank you also goes to my second supervisor, Richard Vipond, who always made time for thesis catch-ups and regularly checked how I was doing, both project progress-wise and personally; always making sure that I (at least tried to) keep a balance. Thank you for supporting this project and me for the whole duration. Miles Carroll, Julian Hiscox and Richard Vipond, this project would not have been there without you all, and to be fair neither would have I. Thank you for entrusting me with this project, supporting me at all the different time-points and always being available when most needed. Thank you to Roger Hewson for all the cool science chats and finding time despite your busy schedule.

Thank you to everyone in the Genomics Team, RIPL, DSP and V&P who assisted whenever I needed help and particularly because they were always happy to do so. Specifically, I am grateful to Kui Lewandowski, Dan Carter, Ant Crook, Sarah Lumley for their help, both scientifically and personally.

The biggest impact of this thesis came from work built on a collaboration and for this I am grateful to Stephan Günther, for trusting in me when I first approached him about initiating this part of the project and for giving me the amazing opportunity to pursue this collaboration. I also owe a huge thank you to Sophie Duraffour, who's passion for her work inspired me in our first encounter and continues to do so to-date; your drive and commitment has always reminded me why it's important to *never give up*.

Thank you to everyone in the BNITM and ISTH team, as everyone's commitment and help brought this work together, it would not have been possible without them. I would like to particularly thank Lisa Oestereich, Anke Thielebein, Meike Pahlmann, David Wozniak, Annick Renevey and Julia Hinzmann who all helped at various time-points and in different aspects with setting up, logistics and the more practical side of things.

A big thank you also to Phillippe Lemey, not only for the involvement in my PhD, but also for providing the next step for my research and above all for his support and understanding during the last year. I am really excited to be part of your research group and looking forward to the opportunities and challenges ahead.

Deborah, my sweet friend, thank you for being there every step of the way and always taking care of everyone and everything you possibly could, particularly me. All the trips to the airport definitely made the travel less bitter and more interesting - especially when the car decides to give up on us! Lizzy and Scottyroddy your combined efforts to encourage me definitely worked, it was lovely to have a short-term flatmate and neighbour around the corner for all those tired evenings when heading back home alone just didn't cut it, thank you for being there even when not in lovely Salisbury anymore! Nadina, you made me feel like part of your lovely family for all the weekends I spent with you, thank you for every delicious meal and your constant care and concern for me. Ant and Dan, Five Guys will never be the same without you and neither will my soft-drink habits; our cinema trips added some spice (or should I say salt) to my long weeks of work and looking back I was fortunate enough to have obtained enough salt to make it through the whole of this project, plus a few good scares to keep me awake during long working days. Slummers, those cheeky half pints kept me sane and despite being far away you made sure to actively keep checking up on me - sanity-check, check! Sophie, the support has not only been scientific, so here's one more thank you for all the non-scientific support you have provided along the way, to the glass always half full - cheers! Anke, not sure I need to say much, but the least I can say is thank you for the constant and much appreciated hugs. Sophie, Anke, Meike, Lisa, David, Jule, Jonas and Ellie, Hamburg would have definitely not been as fun without you all! Magda, don't forget what always keeps us going: at least we don't have to do a PhD in MEMS. Lieseke, thank you for being there from even before the start and for sharing all the frustrations along the way. Dulcinha, thank you for always believing in me and supporting me in the good and bad moments, taking so good care of me at some of the most difficult and frustrating times and being so happy and proud of me at all the good ones. Joao, akiro. To every single person included in this section, everyone that is not specifically mentioned and the ones I might have forgotten, I would like to say a big thank you.

A big thank you goes to Kyriakos Efthymiadis who shared my worries, was always the voice of logic, knows how to reason with me and helped me get through the most frustrating days; your support has been priceless.

Finally, I would like to dedicate this thesis to my parents and my brother for their support through these years, when things were good but mainly when things were not. For understanding and being there for me even when I was stressed out and in the worst of moods and for never stopping to believe in me. To my dad who somehow completely by chance led me exactly where I was meant to be; your advice has always been so valuable. I have always admired your work ethics and love, the way you care for what you do and everything you are passionate about. It is inspiring and mesmerizing to listen to you explain any fact, scientific or not, and your knowledge for everything never stops to amaze me! Your goals, determination and views on life make us all stronger by the day. To my mum, who none of us understand how you manage to juggle everything or manage to deal with all of us, thank you for always listening and always being there when we need you the most. You always make every difficult moment easier and every storm calmer; I really don't know how we would manage without you and not just because you sort out our taxes. To my brother for entertaining me constantly with all the silly internet videos and for his unconventional sibling love; thank you for always accepting me just the way I am and definitely not said very often but I am very proud of you. It has been a tough year but if anything makes it better it's that we all have each other.

It has been an absolutely amazing trip and I have loved every single moment. I complete this thesis with more knowledge as a scientist but more importantly as a better version of myself. If there is anything I could change I wouldn't change a thing, I have loved every second and all the past moments have brought me to where I am now.

Τέλος, μπορεί όσο περνάει ο καιρός να μεγαλώνουμε και να ωριμάζει το μυαλό αλλά στη καρδιά θα παραμένω πάντα παιδί, για την ακρίβεια, ένα μικρό ροδομηλάκι.

Metagenomic next generation sequencing for viral pathogens: Application and validation of a deployable sequencer for virus identification

Liana Eleni Kafetzopoulou

ABSTRACT

Viral pathogen identification and discovery is of significant importance to clinical virology and public health. Rapid and unbiased diagnostic methods are vital when developing a strategy for treatment and eradication of an emerging pathogen. Advances in the field of sequencing over the last decade have facilitated pathogen identification without prior knowledge, introducing its potential and importance as a diagnostic support approach. More recently the development of the MinION, a portable palm-sized sequencing device run via a laptop computer has introduced sequencing in real time in remote settings and overcome many limitations of its predecessors. The objective of this project was to investigate the possibility of non-targeted viral pathogen identification and full genome recovery from clinical samples utilising a metagenomic sequencing approach coupled with the MinION. The limitations and bottlenecks identified during initial experiments using a model system of Hazara virus spiked into human serum and preliminary clinical sample testing allowed for the development and optimisation of a pipeline for clinical sample processing. It was subsequently demonstrated that direct metagenomic sequencing of clinical serum samples in a laboratory setting can elucidate full viral genomes directly from clinical samples for Chikungunya, Dengue and Lassa virus across a range of clinically relevant viral titres. Finally, to evaluate metagenomic nanopore sequencing for the recovery of whole viral genome sequences from clinical samples of a divergent virus in a remote and resource-limited setting, metagenomic nanopore sequencing of Lassa virus was implemented in Nigeria during the 2018 endemic season, providing vital public health information in real-time.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

ATP	Adenosine Triphosphate
BAM	Binary Alignment Map format
BNITM	Bernhard Nocht Institute for Tropical Medicine
BWA	Burrows-Wheeler Aligner
CCHFV	Crimean-Congo Haemorrhagic fever virus
CFR	Case fatality rate
CHIKV	Chikungunya Virus
CNS	Central nervous system
CoV	Coronavirus
Ct	Cycle threshold
DENV	Dengue Virus
DEPC	Diethyl pyrocarbonate
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxynucleotide
dsDNA	Double stranded DNA
dsRNA	Double stranded RNA
EBOV	Ebola virus
EtOH	Ethanol
HAZV	Hazara Virus
ICTV	International Committee on Taxonomy of Viruses
ILFRC	Institute of Lassa Fever Research and Control
ISTH	Irrua Specialist Training Hospital
L segment	Large segment
LASV	Lassa Virus
M. natalensis	*Mastomys natalensis*

MERS	Middle East Respiratory Syndrome
NCDC	Nigeria Centre for Disease Control
NFW	Nuclease Free Water
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
PHE	Public Health England
qRT-PCR	Quantitative Reverse Transcription-PCR
RdRp	RNA-dependent RNA polymerase
RIPL	Rare and Important Pathogens Laboratory
RNA	Ribonucleic acid
RT	Room Temperature
RT-PCR	Reverse Transcription-PCR
S segment	Small segment
SAM	Sequence Alignment Map format
SARS	severe acute respiratory syndrome
SISPA	Sequence Independent Single Primer Amplification
SPIA	Single Primer Isothermal Amplification
ssRNA	Single-stranded RNA
VHF	Viral haemorrhagic fever
WHO	World Health Organisation
WNV	West Nile virus
YVF	Yellow fever virus
ZIKV	Zika virus

Chapter 1

Introduction

1. Chapter 1. Introduction

1.1 Viruses

Viruses are classified into viral taxa and as defined by the International Committee on Taxonomy of Viruses (ICTV) (*1*), "a virus species is a polythetic class of viruses that constitute a replicating lineage and occupy a particular ecological niche". More specifically ICTV specifies that a "polythetic class" is defined by members, which have several properties in common, but not necessarily all or a single uniform one across all members, subsequently members of a viral species are defined by a consensus group of properties. The classification system includes, where possible, the assignment of viruses as members of appropriate species. Viral species are classified as members of recognised genera and some genera are members of recognised sub-families. Viral taxonomic classification is as follows: (Order), Family, (Sub-family), Genus, Species. The second widely used classification system for viruses is the Baltimore classification, first described in 1971 by David Baltimore (*2*). The Baltimore classification categorises viruses based on the type of their genome and their replication strategy. Seven classes of viruses exist in the Baltimore Classification system: (I) double stranded DNA viruses, (II) single stranded DNA viruses, (III) double stranded RNA viruses (dsRNA), (IV) positive-strand single-stranded RNA viruses ((+)ssRNA), (V) negative-strand single-stranded RNA viruses ((-)ssRNA), (VI) positive-sense single-stranded RNA reverse transcriptase viruses and (VII) double stranded DNA reverse transcriptase viruses (*3*).

The majority of emerging and re-emerging infectious diseases are caused by RNA viruses, making RNA viruses a huge public health burden and prominent among the emerging viruses (*4*, *5*). RNA viruses are subdivided into three groups based on the structure and coding of their genome: (a) (+) ssRNA (b) (-) ssRNA (c) dsRNA. Emerging and re-emerging viruses of major public health concern are in their vast majority (+) ssRNA and (-) ssRNA viruses. Genomes of (+) ssRNA viruses are single-stranded molecules of RNA, function as mRNAs and are infectious. The viral genome serves both as the mRNA and as the template for synthesis of additional viral RNAs. Genomes of (-)ssRNA viruses contain a single stranded RNA genome that may be segmented (non-segments or segmented into 2-8 molecules) and the naked genomic RNA is not infectious (*6*). The primary event following genome release within a host cell is transcription, which is successfully accomplished in the presence of the viral proteins. Emerging and re-emerging viruses are extensively researched to

understand the way they infect and cause disease in humans, animals or plants. Additionally, non-pathogenic viruses are very commonly used as they can provide safe alternatives to study their related pathogenic-viruses. An overview is presented in this section for each virus studied in this thesis.

1.1.1 Hazara virus

Hazara virus (HAZV) was first isolated in 1964 from *Ixodes redikorzevi* ticks in Western Pakistan and identified to be a distinct virus but serologically related to Crimean-Congo Haemorrhagic fever virus (CCHFV) (*7*, *8*). HAZV natural mammalian hosts are largely unknown, although antibody evidence has been found in wild rodent sera (*9*, *10*). HAZV is not pathogenic to humans and can be handled in a low containment laboratory (CL2) making it a suitable model virus for molecular experiments to understand its related pathogenic nairovirus (*11*, *12*).

HAZV belongs to the family *Nairoviridae*, the genus *Orthonairovirus* and the order *Bunyavirales (3, 10, 13).* Orthonairovirus genomes are single-stranded, negative-sense RNA and are composed of three segments; the large (L) segment (~6.8-12 kb), the medium (M) segment (~3.2-4.9 kb) and the small (S) segment (~1-3kb). The HAZV genome consists of an L segment of ~11.9kb, an M segment of ~4.5kb and an S segment of ~2kb. All three segments have the same complementary sequence of nucleotides at their 3' and 5' end. These terminal nucleotide sequences are highly conserved within each viral genus.

Orthonairovirus virions are enveloped and have a diameter of ~80-100nm. Glycoproteins are embedded within the lipid bilayer envelope, which has a thickness of 5-7nm, and display projections of 5-10nm (*10*). The virion contains individual complexes of each genomic RNA segment and nucleoprotein, termed ribonucleoprotein particles. Each genomic segment contains a single open reading frame, which is flanked between non-coding regions. The three segments combined encode for four structural proteins (Figure 1.1): the L segment encodes the RNA-dependent RNA polymerase (RdRp), the M segment the two glycoproteins (G_N and G_C) and the S segment the nucleoprotein (N) (*3*). The L protein functions as an RdRp but also has endonuclease activity and is responsible for generating the capped primers required for transcription. The M segment is translated into a single polyprotein precursor, which is co- and post-translationally cleaved into the two glycoproteins. The N protein is the most abundant viral protein and plays an important role in viral replication and protecting viral RNA from degradation.

Attachment of the virion to the host cell surface is mediated by an interaction between the viral glycoproteins and host receptors and leads to viral entry by endocytosis (*3*, *10*). Acidification triggers conformational changes to the envelope glycoproteins, uncoating the virion and leading to fusion of the viral membranes with the endosomal membranes. The viral genome is released and primary transcription of the negative-sense RNA to mRNA is initiated. The virion-associated RdRp mediates transcription by interacting with the three ribonucleocapsids and the generated mRNA is translated to synthesise viral polypeptides. L and S mRNAs are translated by free ribosomes, whereas M mRNAs are translated by membrane-bound ribosomes. Once translated, the glycoprotein precursors undergo post-translational processing in the Golgi membrane. Genomic segments are replicated via the formation of complementary copies of the viral genomic RNA used as a template for the production of subsequent viral genomic RNA (*3*, *10*). Virions assemble following the interaction of the newly synthesised viral ribonucleocapsids with the glycoproteins. The virions bud and are released from the cells through the secretory pathway. At least one of the three ribonucleocapsids needs to be contained within a virion for it to be infective.

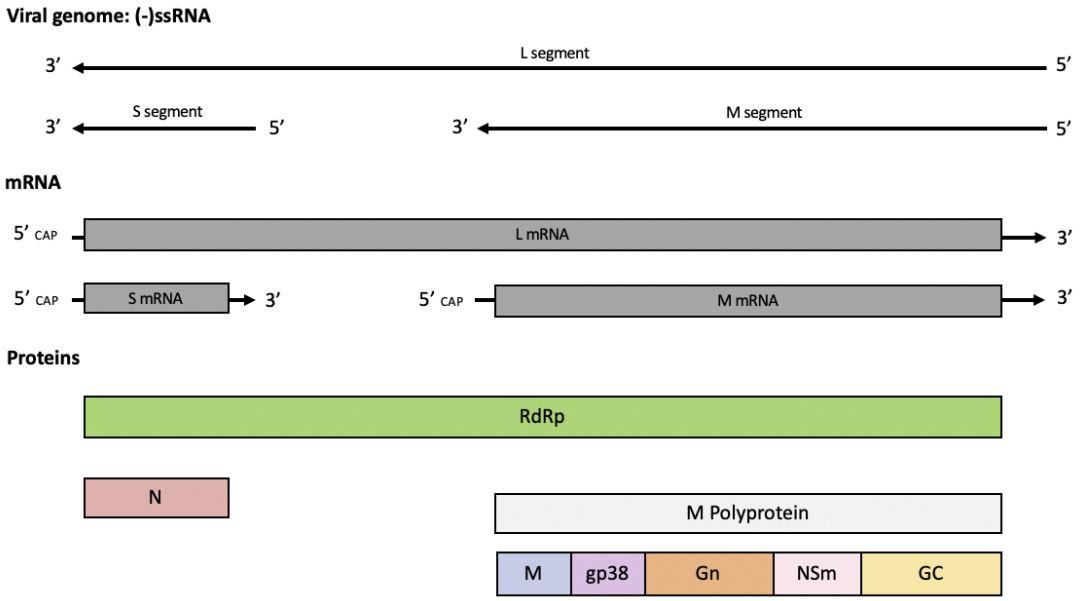


Figure 1.1 HAZV genome structure and protein products.
Each segment contains a single long ORF and a subgenomic mRNA is expressed for each segment. A single polyprotein is produced for the M segment, which is co- and post-translationally cleaved.

1.1.2 Dengue virus

Dengue virus (DENV) is the most prevalent human arboviral pathogen globally (*14*). DENV fever occurs following infection with one of four serotypes (DENV1-4) and clinical presentation typically includes a high fever, arthralgia, myalgia, rash and headache (*15*). A minority of cases develop acute haemorrhagic manifestations and multi-organ failure. Despite DENV cases being underreported a 143.1% increased incidence was estimated between 2005 and 2015 (*16*). According to the World Health Organisation (WHO) approximately 500,000 DENV infected patients require hospitalisation annually with widely varied disease presentation and an unpredictable course of disease (*17*, *18*).

DENV belongs to the family *Flaviviridae* and the genus *Flavivirus (3, 13, 19–21).* Flavivirus genomes are non-segmented, single-stranded, positive-sense RNA and range from 9.5-12.5 kb. The DENV genome is ~11kb in length and is non-polyadenylated with a methyl-G-cap protecting its 5' end and a hairpin structure at the 3'end, which stabilises the genome and is likely involved in genome replication (*3*, *19*, *20*). The viral genome has three distinct roles during the lifecycle of the virus: (i) mRNA for translation (ii) template during RNA replication (iii) genetic material packaged within the new viral particles.

The DENV virion is enveloped with a diameter of ~30nm and has a capsid (*3*, *19*). The glycoproteins lie flat on the surface of the mature virion and the structure is pH sensitive, at low pH the glycoproteins form spikes, which mediate the fusion process. The virions attach to host receptors on the surface of target cells through the viral envelope protein and are endocytosed. The low pH triggers the envelope protein rearrangement and leads to fusion of the viral envelope and cell membrane. Membrane fusion releases the viral RNA, which is efficiently translated utilising the host cell machinery. The positive-sense RNA genome is translated to produce viral proteins in the endoplasmic reticulum and a minus-strand RNA molecule is synthesized and used as a template for viral replication. Generated structural proteins and genomic positive-sense RNA molecules are packaged and assembled into virions. Progeny virion assembly occurs in the endoplasmic reticulum by budding and following maturation in the Golgi apparatus, virions are released by exocytosis via the host secretory pathway.

The viral genome serves as the single mRNA and is translated from a single open reading frame (ORF) (Figure 1.2) (*3*, *19*, *20*). The genome organisation of *Flaviviridae* is conserved with the structural proteins (envelope, capsid) encoded at the 5' and the RdRp at the 3'end of the single ORF. Due to DENV lacking a poly(A)

tail at the 3'end, translation is cap-directed and viral proteins are synthesized as part of a large single polyprotein precursor. The polyprotein precursor is co- and post-translationally cleaved into ten proteins, three structural proteins (capsid, membrane and envelope) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NSB5). A single capsid (C) protein is encoded and the two DENV envelope glycoproteins (prM and E). The non-structural proteins include a glycoprotein important for replication and flavivirus-specific humoral responses (NS1), a membrane-spanning protein involved in replication (NS2), a large multifunctional protein encoding enzymatic activities (the N-terminal domain is a serine protease and the C-terminal domain is an RNA nucleotide triphosphatase helicase (NS3)), proteins important for genome replication (NS4A, NS4B) and the RdRp (NS5B). Many DENV non-structural proteins interfere with the host immune response and have multifunctional roles in the virus's immune evasion. Replication takes place in the cell cytoplasm and involves the synthesis of a genome-length negative sense RNA using the viral genome as a template. The resulting dsRNA serves as a template for the synthesis of the viral genome copies.

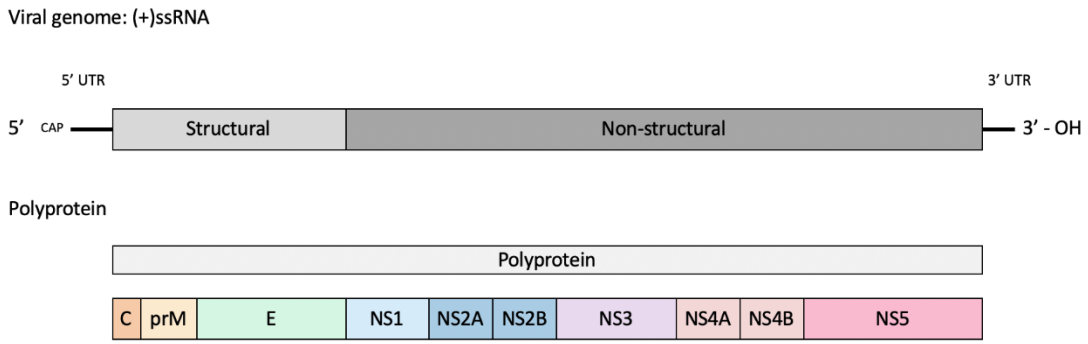


Figure 1.2 DENV genome structure and protein products.
The viral genome contains a 5' methyl cap and lacks a 3' poly-A tail. The genome contains a single long ORF and no subgenomic mRNAs are expressed. A single polyprotein is produced, which is co- and post-translationally cleaved into 11 mature proteins.

DENV can be distinguished in four serotypes, historically differentiated based on neutralisation assay data, DENV-1, DENV-2, DENV3 and DENV-4 (*3*, *19*). The serotypes are related but genetically and antigenically distinct, with 30-40% variation in the viral envelope proteins. Within a given serotype the amino acid homology has a conservation level of ~90% and despite their differences all four serotypes cause similar disease and share epidemiological features.

1.1.3 Chikungunya virus

Chikungunya virus (CHIKV) causes the debilitating arthritic disease, chikungunya fever (*22*). The virus was originally restricted to limited regions of Africa and Asia but has recently spread globally and is designated a serious emerging disease by the WHO (*23*). The majority of people infected are symptomatic and contrary to other arboviral diseases CHIKV presents with less than 15% asymptomatic seroconversion (*24*). Symptoms include high grade fever with sudden onset, rash, headache, myalgia and often long persisting arthralgias (*25*). Over the last 15 years, outbreaks of CHIKV have been associated with increased morbidity and possibly mortality (*26*, *27*).

CHIKV belongs to the family *Togaviridae* and the genus *Alphavirus (3, 13, 26, 28, 29)*. Alphavirus genomes are non-segmented, single-stranded, positive-sense RNA and range from 11-12 kb. The CHIKV genome is ~12 kb in length and is 3'-polyadenylated with a methyl-G-cap at the 5' end (*3*, *28*). Non-structural proteins are translated from the genomic RNA and the structural proteins from the subgenomic RNA. Three types of RNAs are produced by virus-infected cells: (i) genome plus-strand RNA, (ii) complementary minus-strand RNA and (iii) subgenomic mRNA (*28*).

The CHIKV virion is 70nm in diameter and has an envelope containing 80 spikes, each composed of an E1 and E2 transmembrane glycoprotein trimer (*3*, *28*). The protein spikes are contained within a lipid bilayer, which surrounds the capsid and the capsid has a T=4 icosahedron structure, with each subunit assembled from 240 copies of a single capsid protein. Virions attach to host receptors and are endocytosed. Low pH triggers the fusion between the viral envelope and the endosomal membrane, releasing the viral RNA into the cytoplasm. Replication proteins are translated and processed enabling the replication of the viral genomic RNA and the translations of the subgenomic viral mRNA. The structural proteins are produced and cytoplasmic assembly of the genomic RNA and the capsid leads to the formation of the nucleocapsid. The nucleocapsids associate with the processed glycoproteins at the plasma membrane, resulting in budding of the mature progeny virions.

The positive-sense RNA genome serves as an mRNA for the synthesis of the non-structural proteins (nsP1, nsP2, nsP3 and nsP4) and a subgenomic mRNA serves as template for the synthesis of the structural proteins (C, E1, E2, E3 and 6k) (Figure 1.3) (*26*, *29*). Two polyproteins are produced directly from the viral genome: the smaller one (P123), which terminates at an opal codon and the second larger polyprotein (P1234), which includes the gene encoding for the RdRp, which is

produced at a lower frequency (~10-20%) when readthrough of the opal codon occurs. The polyproteins are subsequently processed to generate precursor and end-product non-structural proteins. A short untranslated region located upstream of the structural polyprotein ORF encompasses a subgenomic promoter, which drives the synthesis of a subgenomic mRNA (26S RNA) (*3*, *28*). A single polyprotein is produced from the subgenomic mRNA, which is co- and post- translationally cleaved. The non-structural proteins include a protein encoding methyltransferase and guanylyltransferase activity (nsP1), a protein encoding multifunctional enzymatic activity: nucleotide triphosphatase helicase, RNA triphosphatase and protease (nsP2), a phosphoprotein important for the minus-strand RNA synthesis (nsP3) and the RdRp (nsP4). The structural proteins include a single capsid (C) protein, two envelope glycoproteins (E1 and E2), and two additional products whose functions are not fully understood, the E3 protein and a small peptide (6k). Genome replication occurs in the cell cytoplasm and is initiated with the synthesis of minus-strand RNA. Both minus-strand RNAs (detected at early stages of infection) and plus-strand RNAs are transcribed under the control of non-structural proteins.

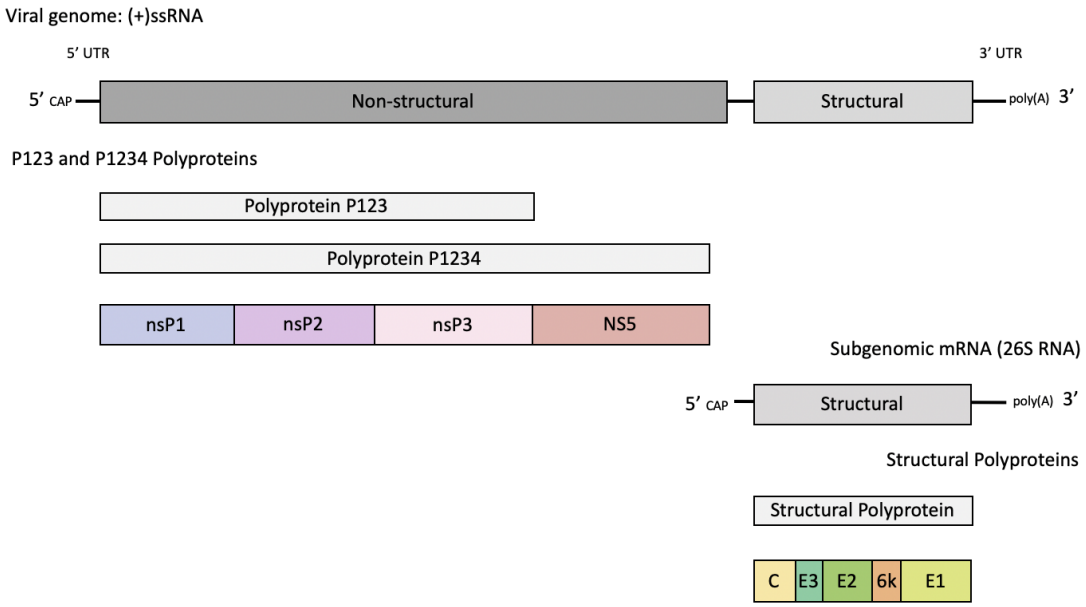


Figure 1.3 CHIKV genome structure and protein products.

The viral genome contains a 5' methyl cap and a 3' poly-A tail. The genome encodes the non-structural proteins at the 5' end, which are synthesized as two polyproteins, a small one and a larger one, which includes the RdRp, both are cleaved by viral proteases. Structural proteins are encoded at the 3' end and synthesized as a polyprotein precursor from a subgenomic mRNA.

CHIKV and DENV are arboviruses of particular interest as they are predominantly transmitted to humans via *Aedes* species mosquitoes, more specifically *Aedes aegypti* and *Aedes albopictus (30, 31)*, and share clinical presentations of arthralgia, myalgia, high fever, headache and rash. Circulation of CHIKV, DENV (and other arboviruses) in overlapping locations leads to challenges in differential diagnosis, especially in endemic regions in which diagnosis is predominantly symptom-based (*32*). Additionally, reports of arboviral coinfections are increasingly common (*33*–*36*).

1.1.4 Lassa virus

Lassa virus (LASV) causes Lassa fever, an acute viral haemorrhagic illness estimated to affect between 300,000 and 500,000 people annually (*37*, *38*) with a typical case fatality rate (CFR) of 18%, although significant increases of up to 31% have been reported during endemic seasons or among hospitalised patients (*39*). Lassa fever was first described in 1969 in the town of Lassa, Nigeria (*40*) and is endemic in parts of West Africa, particularly the countries of Nigeria, Benin, Côte d'Ivoire, Mali, Sierra Leone, Guinea and Liberia (*39*–*44*). Human to human transmission events have also been reported, through direct contact with bodily fluids of an infected patient, primarily as nosocomial infections (*45*–*47*). Laboratory diagnosis poses a significant challenge for LASV due to its high genomic diversity with strain nucleotide variation of up to 32% and 25% for the L and S segments respectively (*48*).

LASV belongs to the family *Arenaviridae* and the genus *Mammarenavirus (3, 13, 49*–*53).* Mammarenavirus genomes are segmented, single-stranded, negative-sense RNA with two segments, a large (L) segment (~7.5 kb) and a small (S) segment (~3.5kb). Both segments are ambisense meaning that each segment contains two open reading frames. The two open reading frames contained within each segment are oppositely oriented and separated by an intergenic region. Viral mRNAs are capped at the 5' end, not polyadenylated and contain a hairpin structure at their 3' end. The capped 5' ends are followed by 4-5 nucleotides, which are not found in the viral genome, acquired through the cap-snatching mechanism involved in the priming of the viral mRNA synthesis (*54*–*58*). Sequences on the 3' end of both segments are highly conserved and base complementarity (19-33 nt) is present between the 5' and 3' ends of each segment, able to form panhandle structures, which modulate replication and transcription (*55*, *59*).

The LASV virion is round, pleomorphic and enveloped with a diameter reported to have a range between 40 and 300 nm (*49*, *55*). The two glycoproteins (GP1 and GP2) and the stable signal peptide (SSP) form a trimer complex, which extends from the virion surface and forms spikes. Within the virion, the genome segments are encapsidated by nucleoprotein (NP) and associated with the L protein (RdRp) forming ribonucleoprotein particles (ribonucleocapsids) organised in circular structures. The virion attachment to the host cell is mediated by the GP1 glycoprotein extracellular spike and binding of the virion to the host cell surface activates endocytosis. Structural changes are triggered by the acidic pH leading to fusion between the viral envelope and the endosomal membrane. Subsequently, transcription and replication of the viral genome take place in the cytoplasm. Ribonucleocapsids associated with Z and glycoprotein complex at the plasma membrane where the progeny virions are released by budding, driven by the Z protein.

The L genome segment encodes for the RdRp and the Z protein. The S genome segment encodes the nucleoprotein (NP) and the glycoprotein complex (GPC) (Figure 1.4) (*3*, *49*). The intergenic regions present in both segments serve as a termination signal during transcription, likely through the formation of secondary structures (hairpin formation) which prompt the release of the polymerase from the substrate (*60*). Following the virion entry into the host cell, the RdRp initiates transcription to produce the NP and L mRNAs. Transcription is initiated at the 5′ untranslated region and terminates at the non-coding intergenic region found on both segments. The RdRp continues past the intergenic region as the concentration in NP increases, creating full length complementary segments (antigenome, (+) ssRNA). The complementary segments are templates for the transcription of GPC mRNA and Z mRNA and serve as a template for the replication of the viral genome. A glycoprotein precursor polyprotein is encoded on the S segment, which is post-translationally modified by a signal peptidase and an ER protease. The signal peptidase cleaves the N-terminal signal sequence of the polypeptide to produce a stable signal peptide, which is required for the GP maturation. The remaining GP polypeptide is glycosylated and cleaved by an ER protease into the two glycoprotein subunits, GP1 and GP2.

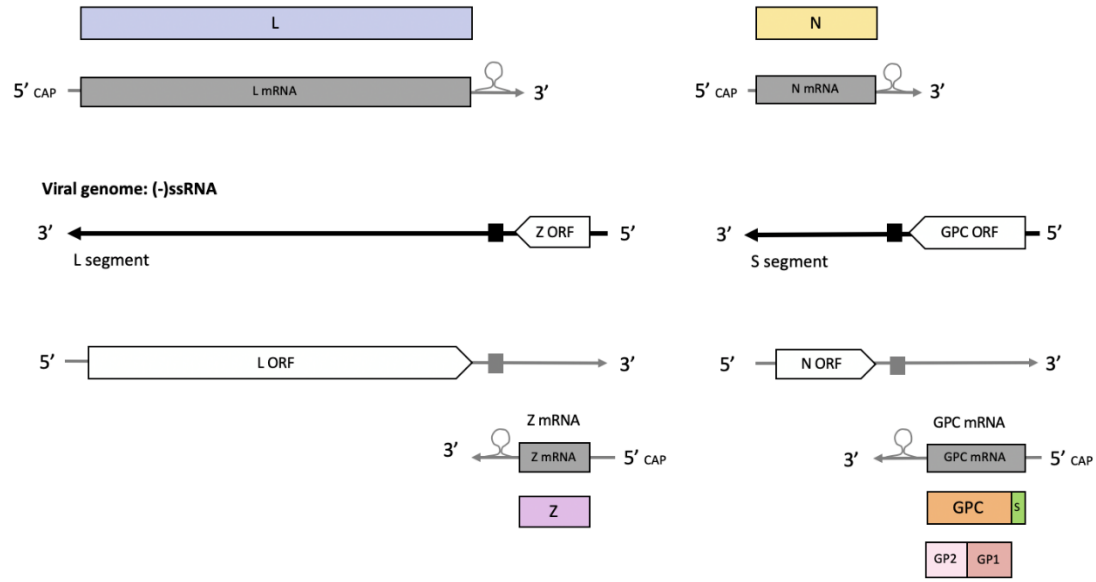


Figure 1.4 LASV genome structure and protein products.

Both segments are ambisense. Each segment contains two open reading frames. The two open reading frames contained within each segment are oppositely oriented and separated by an intergenic region.

1.2 Emerging and re-emerging viruses

Emerging pathogens are defined as the causative agents of an infectious disease that has newly appeared in a population (e.g. Dengue fever: global spread of all four DENV serotypes (*61*)) or has existed and is subsequently rapidly increasing in incidence, prevalence and geographic distribution (*62*, *63*) (e.g. Ebola haemorrhagic fever: 2014-2016 West Africa Ebola virus (EBOV) outbreak (*64*)). Re-emerging pathogens are characterised by resurgence, after a substantial decline, of a disease that previously caused major public health problems globally or in a particular country (*65*) (e.g. Yellow fever (YFV): West Africa resurgence since 2000 (*66*)). According to the WHO, both emerging and re-emerging diseases as defined above, can be described as emerging, as it defines an emerging infectious disease as "one that has appeared in a population for the first time, or that may have existed previously but is rapidly increasing in incidence or geographic range" (*67*).

The recent global emergence of zoonotic viruses has caused significant human disease. A combination of factors including globalisation, environmental changes, human intervention and their effect on disease spread and vector distribution have led to new opportunities for viral pathogens, leading to both the emergence of previously unknown viruses and the re-emergence of viruses previously not thought to be of public health significance (*68*–*70*). The majority of these infectious viral diseases are caused by RNA viruses (*5*, *71*) which emerge either because they acquire the ability to infect new host populations that were previously non-susceptible (*63*, *72*) or because of changes in their geographical spread introducing them to new, previously unaffected, populations (*73*). The vast majority of viral pathogens are zoonotic, originating in the animal world and following incidents of spillover to human hosts, they adapt and subsequently spread to cause an epidemic or in certain situations a pandemic (*70*). Viral pathogens that have developed the ability to cross species barriers can cause devastating infections leading to a high morbidity, mortality and disease burden (*74*). Many of these zoonotic viruses can cause little to no disease in their natural hosts, and interspecies transmission to humans may be asymptomatic, but can also lead to disease with symptoms ranging from mild to severe manifestations (*75*).

1.2.1 Zoonotic viruses

Over 70% of emerging infectious diseases affecting the human population are known to be zoonotic, infectious diseases transmitted from an animal host to humans (*76*). Human infection occurs through two routes (a) pathogens spread through direct or indirect contact with infected animals and (b) pathogens transmitted through vectors (*77*).

Direct contact transmission requires physical contact between an infected animal and a susceptible human. Most commonly the transfer of the viral pathogen occurs through saliva (e.g. bites or scratches). Lyssavirus is the most prominent among the direct transmission zoonotic viruses, spanning across all continents except Antarctica (*78*, *79*). It has been identified through historical documentation to date back at least 4000 years and is mainly transmitted through bites from rabid animals causing a lethal encephalitis; Rabies disease (*79*). Zoonotic indirect contact transmission occurs through consumption of contaminated food (e.g. human consumption of infected bush meat) and water or by close contact with faeces, organs, blood or other bodily fluids of infected animals. LASV is an example of indirect contact transmission as it is spread through human exposure to urine or faeces from the infected host rodents *Mastomys natalensis*, in which the virus causes persistent asymptomatic infection (*80*, *81*).

Finally, transmission can occur via intermediate species (vectors) which carry and transfer the virus, facilitating its spread. Climate and habitat changes have had a tremendous impact on vector-borne zoonotic viruses, introducing new geographical ranges to the vectors and expanding the population of susceptible animals and humans exposed to the virus (*82*). An example of a vector-borne zoonotic virus is West Nile virus (WNV), which originated in Africa and has spread to Asia, Europe and the Americas causing periodic large outbreaks (*83*). It is maintained in an enzootic cycle involving vertebrate hosts (reservoir) and mosquitoes (vector) (*84*). The vast majority of arthropod-borne viruses (arboviruses), such as WNV, are zoonotic and of major importance due to their increased frequency in recent years leading to a significant global health burden (*85*). The term “arboviruses” encompasses all viruses, which require hematophagous (blood-sucking) arthropods such as mosquitoes and other biting flies or ticks to maintain their transmission cycle (*86*). Transmission can occur through a sylvatic or enzootic cycle in which the pathogen circulates among wild animals and can cause disease in incidental or dead-end hosts such as humans, following events of spillover events (*87*). Epidemics caused by arboviruses have increased in prevalence over recent decades, with the spread of mosquito-borne

arboviruses such as YFV, DENV, WNV, CHIKV and Zika virus (ZIKV) across both hemispheres (*88*). The significance of arboviruses is highlighted by DENV, CHIKV and YVF, which are arboviruses of immense global health burden as they have lost the need for enzootic amplification, expanding their host range to include humans as an amplifying host and consequently be responsible for extensive epidemics (*87*, *89*).

1.2.2 Disease manifestation

Emerging viruses present an important challenge to public health as effective interventions are required to control outbreaks and eliminate global health and economic risks linked with such an extensive disease burden (*90*). Most emerging viral infections share common clinical features with mild/moderate disease presentation of fever and flu like symptoms, sometimes associated with rash (CHIKV, ZIKV), arthralgia (CHIKV, Mayaro virus) or jaundice (YFV, Rift Valley fever virus) and more severe disease manifestations with encephalitis or menencephalitis (Nipah virus, YFV, CHIKV, ZIKV) and haemorrhagic fever (LASV, Rift Valley fever virus, CCHF) (*91*).

Viral encephalitis is a type of neuroinvasive disease caused by a viral infection and refers to an acute, commonly diffuse inflammatory process affecting the brain parenchyma (*92*, *93*). Neurotropic viruses invade the central nervous system (CNS), causing neuroinvasive disease and come from a variety of different families, including *Flaviviridae* (e.g. WNV), *Paramyxoviridae* (e.g. Nipah virus), *Togaviridae* (e.g. Western equine encephalitis virus), *Arenaviridae* (e.g. Junin mammarenavirus) and *Bunyaviridae* (La crosse virus). Clinical presentation of encephalitis in patients may include change in consciousness level, fever, seizures, movement disorder or focal neurological defects, and the burden of the disease is big, as many patients with encephalitis face life-long residual physical or neurophysiological defects which require long-term management (*94*). An example is tick-borne encephalitis virus, which, as its name suggests, causes a human disease involving the CNS and occurring in parts of Asia and Europe. The virus is transmitted by infected ticks and the far eastern subtype is associated with the most severe disease presentation with mortality rates of up to 35% (*95*).

Viral haemorrhagic fever (VHF) is a term used to refer to severe illness sometimes associated with bleeding (*96*). VHFs are caused by RNA viruses belonging to the families of *Arenaviridae* (e.g. LASV), *Bunyaviridae* (e.g. Rift Valley Fever), *Filoviridae* (e.g. EBOV) and *Flaviviridae* (e.g. DENV) (*97*). All haemorrhagic fever viruses are negative or positive-stranded RNA viruses of between one and three

genome segments. The initial presentation of VHFs in patients includes an acute febrile illness which is characterised by fever, musculoskeletal pain, nausea, vomiting and vascular permeability accompanied by conjunctival injection, flushing and petechial haemorrhages (*98*). In more severe cases of the disease bleeding manifestations may present accompanied with hypotension, circulatory collapse and shock. The case fatality rate for VHFs has been estimated to reach 80% of severe cases, with the severity of disease depending on the virus, its inoculum and route of exposure (*98*). An example of a VHF is Lassa fever (*99*).

1.3 Virus discovery, diagnosis and genomics

The revolutionary developments of sequencing in the 1970s and of PCR in the 1980s, along with the advances in sequencing technologies over the past 20 years (details in Section 2. History of sequencing) played a critical role in the field of virology and its progress. The technological developments witnessed in the first two decades of the 21st century have transformed viral research from lengthy protocols and retrospective investigations to near real-time pathogen investigation and genomics informed epidemiology. Whole genome sequencing has been instrumental in the effective management of viral pathogens, allowing for pathogen discovery, accurate diagnostic method development, and pathogen surveillance and molecular epidemiology. Sequencing has become a critical tool in clinical microbiology, where it is currently extensively used to facilitate the detection and monitoring of pathogen outbreak and epidemics using genomic precision which was not previously possibly using epidemiological data alone (*100–103*)**.** Sequencing advances have had an immense effect on viral outbreaks both in human and animal populations allowing for efficient identification of the causative agent and for public health surveillance and epidemiological studies, which have in the past been hindered by lack of data.

1.3.1 Virus discovery

The first viruses, described as filterable agents, were discovered as early as the 1890s. In 1886 Adolf Mayer was the first to describe tobacco mosaic disease and demonstrate its infectious nature. Mayer successfully infected healthy plants by inoculating them with fluid extracted from diseased plants. He attempted to identify the infectious agent but could not find any microorganisms accountable, leading to his observation that the soluble agent was enzyme-like. He concluded that the cause of the disease must be bacterial but the infectious forms had not yet been isolated (*104*). In 1892 Ivanofsky reported that tobacco mosaic infected leaf extracts were still infectious following filtration with filters known to retain bacteria, the first demonstration of a filterable pathogen. In 1898 Beijerinck observed that the filterable pathogen could diffuse through bacteria retaining agar and that it could only be cultured in living, growing plants. Beijerinck described the filterable pathogen as "contagium vivum fluidum" (Latin: "contagious living fluid") and was the first to use the term virus (*105*). The three aforementioned scientists contributed to the discovery and formation of a new concept: the virus, an infectious agent smaller than bacteria, not visible by light microscopy and which only multiplied in the presence of living cells (*3*).

Shortly after, Loeffler and Frosch identified a filterable agent as the cause of foot-and-mouth disease, leading to the description of the first animal virus, foot-and-mouth disease virus (*106*).

The first human virus discovered was YFV. Reports of the yellow fever disease date back to 1498 but it was not until the 18th century that its public health importance was realised and only at the start of the 20th century that the disease began to be understood (*107*). The disease is spread through a mosquito vector, Aedes aegypti, a hypothesis originally formed by Carlos Finely in 1881 and later confirmed by Walter Reed and his colleagues in 1901 (*108*). James Carroll, who was part of Reed's team, showed that serum from a yellow fever patient was still infectious after filtering and therefore was caused by a filterable agent; a virus (*109*).

Viral discovery has been primarily driven by the identification of diseases and the need to understand their causes. Multiple scientific developments over the years have provided essential tools for the investigation of viral pathogens, including: porcelain filters, light and electron microscopy, animal models, tissue and cell culture, immunoassays, PCR and microarrays (*110*). In recent years technological advances and progress in the field of molecular biology have played a vital role in the ability to identify and characterise viral pathogens, particularly sequencing technologies.

Sequencing methods have enabled the detection and identification of viruses in clinical samples, and when used in combination with non-targeted methods, such as metagenomic sequencing make it possible to identify infectious diseases without prior knowledge of the causative agent (*111*). In recent years metagenomics has been used to identify the causative agents of several outbreaks, including Lujo virus in South Africa (21), a novel EBOV, Bundibugyo ebolavirus, in Uganda (22) and novel viruses such as the Rhabdovirus associated with haemorrhagic fever in central Africa (23). Applications of sequencing in virology not only provide information about the identity and the genome of the virus but can also allow for the study of virus/host interactions and host genetic responses in correlation to the infectious agent providing information on the mechanisms of infection and host genetics (*112*).

1.3.2 Virus diagnosis and detection

The most common techniques currently used in diagnostic laboratories for identification of the causative agent of disease include both direct (PCR, antigen detection) and indirect (IgM and IgG serology) methods, but all require prior knowledge of the viral pathogen of interest for their design (*113*, *114*).

RNA viral pathogens are most commonly detected in diagnostic laboratories and acute-care settings by Reverse Transcriptase PCR (RT-PCR) based methods, primarily a major feature due to their timely and accurate nature (*115*). A significant advancement in PCR diagnostics was the development of real-time RT-PCR which enables reliable measurement and detection during amplification process, in addition to the components for the classical/end-point PCR, a fluorescent probe, that emits signal upon binding to target sequence DNA, is used and its signal strength is directly proportional to the number of amplified molecules (*116*, *117*). The combination of the amplification and detection steps in one reaction eliminates the need for post-PCR processing, subsequently reducing the time to result (*117*). PCR methods have high specificity, reliability, reproducibility and can include multiple targets in one assay. Their wide dynamic range allows for analysis of samples with highly variable target abundance and methods are well established in clinical diagnostics with extensive validation, established analysis and reporting guidelines (*117*). The limitations of PCR based methods include their high sensitivity, which can lead to false-positive results from background contamination, particularly when handling high sample volumes, or false-negative results due to PCR inhibitory components present in samples. Additional limitations arise from PCR specificity, which leads to inadequate sensitivity when used for highly variable pathogens such as certain RNA viruses (eg. LASV) and it can only be used to identify the presence or absence of the target sequence against which the assay was designed, leading for example to difficulties in designing assays that can easily distinguish between serotypes (eg. DENV 1-4) (*115*). Reverse transcription of RNA and amplification of short fragments of DNA *in vitro* relies on *a priori* knowledge of the viral target sequence (*115*).

Serological methods are routinely used in diagnostic laboratories and can identify if a patient has been exposed to a pathogen by detecting an antibody response. Serological assays can prove useful when detection of the viral antigen or nucleic acid is unsuccessful due to testing toward the end of or after viremia (*114*). The time frame relative to pathogen exposure is a limiting factor for serology testing early-on during infection if an antibody response has not been mounted (*114*, *118*). Serology requires comparison of acute and convalescent antibody titers and has

issues of cross-reactivity of antibodies triggered by viruses within the same family or vaccination. This is the case with anti-flavivirus antibodies which are known to cross react with multiple flaviviruses leading to non-specific diagnosis and inability to distinguish between viral serotypes (*118*, *119*).

1.3.3 Genomic surveillance and epidemiology

Viral disease surveillance involves the continuous monitoring and systematic collection, analysis and dissemination of information relating to the occurrence of a disease within a population (*120*, *121*). The ultimate goal of viral disease surveillance is to facilitate the prevention of disease and effectively control outbreaks (*120*, *121*). Genomic surveillance involves sequencing and analysis of viral genomes collected from patient and animal samples for the purpose of understanding viral spread and the changes occurring during the course of the disease spread within a population.

Viral disease epidemiology is the study of disease events within a population (human or animal) and the detailed description of their incidence and rate at which they occur (*120*). The impact of viral transmission and spread is investigated through the study of different factors such as prevalence, mode or method of transmission, duration of infection, population density, patterns or associations, susceptibility and transmissibility. Viral genomic epidemiology is the study of viral epidemiological and emergence dynamics using sequencing data and genomics associated analysis (*103*).

Evolutionary relationships between and among species were studied as early as the 18th century, although efforts at the time to categorise species were based solely on morphological characteristics (*122*, *123*). Over the course of time, phylogenetic analysis has become a powerful tool in virus epidemiology research. By obtaining molecular/genomic data and applying statistical methods, it is now possible to understand viral emergence and transmission, as well as determine the key factors for successful epidemic spread. Phylogenetic trees are the visualisation of phylogenetic analysis and represent the patterns of shared mutations between genomic sequences as a function of their topological distance (*124*).

The International Health Regulations (2005) as stated by the WHO "require the rapid detection of public health risks, as well as the prompt risk assessment, notification, and response to these risks" (*125*). Genomic investigations have had a profound impact in our understanding of viral diseases and the dynamics of infection (*126*). The 21st century has witnessed a huge leap in sequencing technologies and has introduced molecular precision to the detection and monitoring of viral outbreaks.

Rapid, inexpensive sequencing has paved the way for viral genomics and created large and continuously growing databases (e.g. GenBank) of nucleotide sequences. The plethora of available viral genomes has permitted studies on pathogen diversity, evolution and transmission creating opportunities for genomic surveillance and epidemiology using molecular precision (*103*). Dudas and Bedford in 2019 (*124*) compared the ability of single genes versus full genomes to resolve phylogenetic, temporal and spatial inference. Their study importantly highlighted the ability of whole genome sequences to significantly enhance resolution and impact our understanding of emerging viral pathogens, which is increasingly evident with molecular surveillance and epidemiology shifting from retrospective studies to near real-time investigations (*100*–*103*). Particularly so for RNA viruses, whose polymerases have low replication fidelity leading to a rapid rate of mutations and subsequently a rapid differentiation of viral lineages at the whole genome molecular level (*124*, *127*, *128*). The high mutation and replication rates lead to genetic variations, which can be monitored during their spread (*103*).

One of the first viral challenges of the 21st century was the 2002-2003 Severe acute respiratory syndrome (SARS) epidemic, which affected 23 countries (*129*). Within a time frame of six months the virus was isolated and sequenced allowing for the design of molecular assays, which confirmed that the new CoV was responsible for the SARS epidemic (*130*–*132*).

In early April 2009 a new H1N1 influenza virus emerged and spread rapidly across the world. The first complete genome sequences were released to a publicly available influenza database by the Centre for Disease Control and Prevention (CDC) a few weeks later (*133*). Within a few months early assessments of transmissibility and severity were made using molecular epidemiology. Fraser *et al.* (*134*) identified the transmissibility to be substantially higher than that of the seasonal flu and estimated the basic reproductive number (R(0)) to be comparable to previous influenza pandemics. They also estimated the start of the outbreak to be a few months prior to the first documented case. Subsequent analysis by Rambaut and Holmes (*135*) used a larger data set to update the estimates of rate of evolutionary change and date of origin of the virus. They confirmed the date of origin of the outbreak to be a few months prior to the first documented case and obtained similar population growth rates and epidemic doubling times. In 2006 Mena *et al.* (*136*) conducted a retrospective study and identified H1N1 precursor viruses in central Mexican swine. The authors presented strong evidence that the first human outbreak of H1N1 occurred in early 2009 in Mexico and highlighted the importance of surveillance as

the outbreak rose from a region, which was not previously considered a pandemic risk.

Another viral outbreak that raised similar questions over its origin and transmission is Middle East Respiratory Syndrome coronavirus (MERS-CoV) which has been repeatedly reported since 2012. Preliminary data published by the WHO MERS-CoV Research Group (*137*) suggested the involvement of dromedary camels and concluded that sustained human-to-human transmission was not observed. Haagmans *et al.* (*138*) investigated the presence of the virus in dromedary camels linked to two human cases reported in 2013. Using PCR testing and sequencing they confirmed the presence of MERS-CoV in camels but could not conclusively prove that the camels were the source of infections in humans. Subsequent sequencing studies demonstrated multiple independent introductions into the human population from close contact with infected camels (*139*, *140*). A retrospective study in 2018 by Dudas *et al.* provided a comprehensive genomic investigation of MERS-CoV phylodynamics in humans and camels. The authors confirmed that the viral population is maintained exclusively in camels and that humans are incidental hosts. Additionally, they estimated mean R (0) values of <0.90 suggesting that the virus was not likely to become endemic in humans, however they highlighted the need for continuous surveillance to monitor for the possible emergence of variants with increased transmissibility.

The 2013-2016 EBOV epidemic in West Africa was of unprecedented scale with ~28,500 reported cases and ~11,000 reported deaths (*141*). The epidemic transformed the concept of viral epidemiology and marked the beginning of large-scale phylodynamics analysis to inform ongoing outbreaks (*141*, *142*). Genomic investigations enabled viral genomic surveillance during the unfolding outbreak and assisted in our understanding of the origin, transmission and evolution of the virus. The most recent common ancestor of all sequenced EBOV genomes was estimated to be at the beginning of 2014 which was consistent with classical epidemiological investigations that placed the first case in Guinea around late December 2013 (*143*–*146*). Multiple introductions from the animal reservoir were excluded and it was demonstrated that the viral spread was maintained by human-to-human transmission (*142*–*151*). As the outbreak unravelled there was an apparent shift towards rapid in-country sequencing which assisted in understanding viral transmission chains and community spread (*142*, *147*, *152*, *153*). Most notably the set-up of portable MinION sequencing in Guinea by Quick *et al.* (*147*) to enhance local sequencing capacity and EBOV genomic surveillance.

In 2016, local transmission of ZIKV was identified in 33 countries and ZIKV infection during pregnancy was suspected to cause microcephaly and congenital abnormalities (*154*). Subsequently the 2016 ZIKV outbreak was declared by the WHO as a Public Health Emergency of International Concern. Efforts to gain insights into the origin, transmission and diversity of the virus were initiated and included sequencing of retrospective and ongoing cases (*155*, *156*). Preliminary analysis by Faria *et al.* (*157*) of a small number of ZIKV sequences indicated a single introduction of ZIKV into the Americas. Using molecular clock analysis, they estimated the introduction to have occurred between May and December 2013, more than a year prior to the detection of ZIKV in Brazil. The spread of ZIKV was extensively documented in subsequent studies using genomic epidemiology (*156*, *158*, *159*). Grubaugh *et al. (159)* used genomic epidemiology to investigate ZIKV transmission in Florida and identified that multiple introductions of the virus contributed to the Florida outbreak and estimated that local transmission most likely started several months prior to initial detection, in spring 2016.

The shift from retrospective studies to rapid near real-time investigations is apparent over the past 20 years. Continuous advances in genomic sequencing and phylogenetic analysis have expanded the key questions addressed by molecular epidemiology and increased the use of genomic epidemiology to inform public health response (*103*, *141*, *160*).

1.4. History of Sequencing

The x-ray crystallography image of structure B of DNA (Photo 51, Rosalin Franklin and Raymond Gosling) (*161*), the description of the three dimensional structure of DNA (*161*, *162*), the discovery of DNA polymerase (*163*) and RNA polymerase (*164*), and the cracking of the genetic code are key scientific landmarks in DNA molecular biology that were precursors to the field of nucleic acid sequencing. Initial techniques were only able to measure nucleotide composition, rather than directly determine their order. Then in 1965 Holley *et al.* used pancreatic ribonuclease and takadiastase ribonuclease T1 to selectively cleave an alanine transfer RNA isolated from yeast (*165*, *166*). The fully and partially degraded RNA fragments generated were identified via chromatography and electrophoresis and the order of the original nucleotide sequence was reconstructed to produce the first ever nucleic acid sequence (*165*, *167*).

Within the same year Fred Sanger and colleagues described the two-dimensional (2D) fractionation of ribonuclease digested radiolabelled RNA. They used the unique pattern created, dictated by size and sequence composition, to determine ribosomal and transfer RNA sequences (*168*). The separation and visualization of the RNA fragments in 2D, 'fingerprinting' approach, was used in 1972 by Jou *et al.* to determine the first complete sequence of a protein-coding gene (the bacteriophage MS2 coat protein) (*169*) and later in 1976 by Fiers *et al.* to sequence the last remaining MS2 gene (encoding the replicase) (*170*) marking the completion of the first ever whole genome sequence.

The first DNA fragment sequencing was by Wu and Kaiser (*171*). The cohesive ends of bacteriophage lambda were used as a primer-template for DNA polymerase to incorporate radiolabelled nucleotides. By measuring the number and order of nucleotides incorporated, they identified the nucleotide length and sequence. This principle was soon expanded to include the use of defined oligonucleotides sequences as primers for DNA polymerase to initiate activity at selected regions of any DNA molecule (*172*–*175*). Despite these advances, attempts to determine the sequence of DNA were limited to short sequences and restricted by the lengthy and cumbersome nature of the methods used.

The 'plus and minus' method described by Sanger and Coulson in 1975 was the first to promise accelerated DNA nucleotide sequence determination (*176*, *177*). Oligonucleotide sequences were designed to match a target region of template DNA. These primers directed DNA polymerase synthesis of complimentary copies, of

assorted length, in the presence of the four nucleotides, one of which was radiolabelled. The DNA generated was purified and incubation with DNA polymerase was repeated under 'plus and minus' conditions. The 'minus' system consisted of four separate reactions in which synthesis occurred in the presence of only three nucleotides, leading to products terminating at the position of the missing nucleotide. The 'plus' system utilised the 3' exonuclease activity of T4 DNA polymerase in four separate reactions, with a single nucleotide present in each reaction. In each reaction, the enzyme degraded the DNA molecule until a position containing the reaction-specific nucleotide is reached in the sequence; generating fragments, which have that nucleotide at their 3' end. The 'minus' and 'plus' mixtures were run in parallel on an acrylamide gel and the sequence was extrapolated from the bands present, with the two approaches confirming and complementing each other. The method was used to complete the first ever DNA whole genome sequence, that of bacteriophage ΦX174 (PhiX), in 1977 (*178*).

1.4.1 First sequencing methods

Two key DNA sequencing method advancements were made in 1977: the Maxam and Gilbert chemical breakage method and the Sanger chain-termination method (*179*, *180*). The Maxam and Gilbert (*179*) method involved radioisotope end-labelling of DNA followed by nucleotide-specific chemical cleavage. The DNA molecule was preferentially damaged using conditions that targeted single nucleotides in each separate reaction (A>G, G>A, C, C+T) and the damaged base was subsequently removed from the sugar. Each reaction generated radiolabelled fragments extending from the radiolabelled end to the position of the base cleaved. The products of the four reactions were then separated by electrophoresis using an acrylamide gel and the sequence was deduced from the pattern of the radioactive bands.

Later that year Sanger et al. (*180*) described the chain-termination method, which became the most widely used sequencing method for decades. The method is formally known as the chain-termination method or dideoxy sequencing but has become commonly known as 'Sanger sequencing'. Despite its similarity to its predecessor, the 'plus and minus' method, this approach was significantly faster and more accurate. The principle of Sanger sequencing was based on the observation of Atkinson *et al.* (*181*) that the absence of a 3' hydroxyl group in the dideoxy nucleotide analog results in the termination of extension at the location where it is incorporated during synthesis. In the original Sanger sequencing method, the DNA template was

incubated with primer, DNA polymerase, all four nucleotides (one of which was radiolabelled), and one dideoxy analog. The reaction resulted in a mixture of radiolabelled DNA molecules all of which started at the same 5' but resulted in various lengths, with the dideoxy nucleotide analog at the 3' end of each. A separate reaction for each nucleotide and its chain-terminating analog was set-up leading to a total of four reactions. The reaction products were run in adjacent lanes on an acrylamide gel and the sequence was read from the pattern of bands obtained. Sanger sequencing was a major breakthrough and became the most widely used DNA sequencing method for many years. A timeline including the key scientific advancements mentioned, from Photo 51 (1952) up until the official initiation of the Human Genome project (1990) can be found in Figure 1.5.

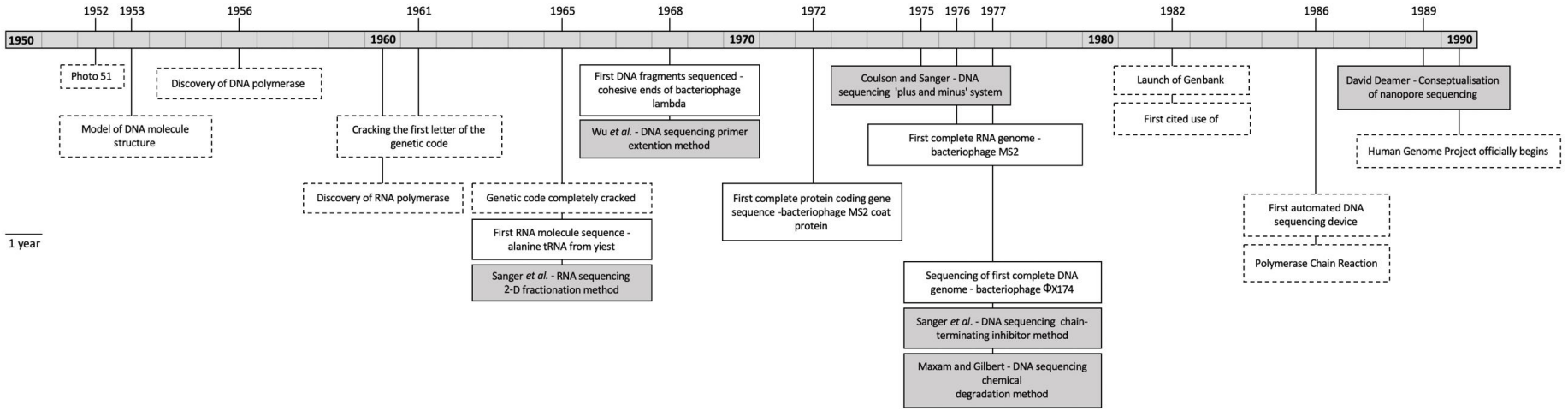


Figure 1.5 Timeline of DNA sequencing relevant advancements up until Sanger sequencing in 1977.

Dashed boxes are events, which were key in the history leading up or driving the advancement of DNA sequencing. White boxes include important sequencing advances and grey boxes important methodological advancements made.

1.4.2 Towards next generation sequencing

The 1980s brought significant progress across the field of molecular biology and in the advancement of DNA sequencing. The launch of the Genbank public database (*182*) and the first use of microarrays in 1982 (*183*), the development of PCR in 1986 (*184*) along with the improved production of high-quality enzymes and primers were just some of the advancements made which proved instrumental for the sequencing field. The development of the first automated DNA sequencing machine and the switch to fluorescent chain-terminating nucleotide analogs in 1986 (*185*) transformed the accessibility of sequencing and enabled the creation of the Human Genome Project, which was initiated in 1990 and delivered the first human genome draft in 2001 (*186*, *187*).

The high demand for simple, rapid and automated sequencing methods led to the next major methodological advancement, which came in 1996 when Ronaghi *et al.* described pyrosequencing (*188*). Pyrosequencing is a sequencing by synthesis approach, which monitors the DNA polymerase activity through the detection of PP_i released during each base addition of the synthesis. In the initial version of the method, a DNA molecule was immobilised on a solid surface and denatured to obtain single stranded DNA. A primer was added and in the presence of DNA polymerase repeated cycles of deoxynucleotide (dNTP) incubation and washing were performed. The second strand synthesis was accompanied by the release of PP_i, which was converted to adenosine triphosphate (ATP) by ATP sulfurylase. The ATP acted as a substrate for firefly luciferase, which catalysed a light-emitting reaction, the conversion of luciferin to oxyluciferin, generating a light signal proportional to the number of nucleotides incorporated during the current round of addition. In contrast to Sanger sequencing, pyrosequencing overcame the need for electrophoresis by combining solid-phase technology with a sensitive luminometric assay and provided a method suitable for processing multiple samples in parallel. Improvements to the method over the following years included the use of paramagnetic beads for the DNA attachment and the use of a nucleotide-degrading enzyme (apyrase) to replace the intermediate washing steps (*189*).

Biotechnology company 454 Life Sciences (later 454 Roche) bought the exclusive rights to pyrosequencing in 2003 (*190*, *191*) and launched the first commercially available next generation sequencing (NGS) platform in 2005 (*192*). The 454 sequencer was a scalable, high capacity sequencing system that used an optimised pyrosequencing protocol. The device consisted of a fluidic assembly, a

chamber where the fibreoptic slide was located and a camera-based imaging system connected to a computer. DNA was randomly fragmented and sequencing specific adaptors added to the fragments. Using limiting dilution, single DNA molecules were obtained and each fragment was bound to individual beads. The fragment captured on each bead was clonally amplified within droplets of emulsion and the amplified template-bound beads were loaded into the individual wells of the fibreoptic slide. Sequencing occurred simultaneously in open wells of the fibreoptic slide and the light generated after every nucleotide addition was measured and signal-processed.

In parallel, Solexa sequencing (later Illumina sequencing), was being developed. The method was conceptualised in 1997 by Shankar Balasubramanian and David Klenerman and in its early form included the immobilisation of discrete single stranded DNA molecules to a surface in order to monitor the incorporation of fluorescently labelled nucleotides by DNA polymerase (*193*, *194*). An advancement that significantly increased throughput and accuracy was the inclusion of molecular cloning, an amplification step which created clonal copies of a DNA fragment within the same location of the sequencing chip, a technology acquired by Solexa in 2004. Molecular cloning was successfully achieved through solid phase bridge amplification, a method previously described by Adessi *et al.* in 2000 (*195*). Illumina sequencing in its current form entails clonal array formation and sequencing by synthesis using a cyclic reversible terminator method for the generation of short-read data (*196*, *197*). Single stranded DNA molecules are immobilized sparsely on the sequencing flow cell prior to solid-phase amplification. Unlabelled nucleotides along with the enzyme are added to initiate bridge amplification, which generates up to 1000 identical copies of each molecule creating a cluster of sequences originating from the same template. Sequencing begins with the addition of four fluorescently labelled reversible terminators, primers and DNA polymerase. A single terminator is incorporated per cluster and, following laser excitation, the fluorescence from each cluster is measured and the incorporated dNTP is identified. The sequencing cycles are repeated and the signal intensities measured are directly basecalled in each cycle. The first commercial sequencer released by Solexa was the Genome Analyzer in 2006 with the capacity to sequence 1 Gb in a single run. Illumina acquired Solexa soon after in 2007 and continued to introduce improvements leading to leaps in sequencing output, with their latest high-throughput sequencer (NovaSeq 6000) released in 2017 capable of producing up to 3000 Gb from a single flow cell.

Pyrosequencing and clonal reversible terminator sequencing were both major methodological advances for the field of sequencing in the late 90s. In the coming

years, a variety of companies developed methods and platforms in parallel, with different sequencing chemistries, read lengths and throughput capabilities (*198*).

Complete Genomics, founded in 2006 and acquired by BGI-Shenzhen in 2013, is a sequencing service focused on whole human genome sequencing (*199*). The company uses sequencing technology that achieves DNA template enrichment in solution through rolling circle amplification leading to single stranded DNA palindrome-promoted coils known as DNA nanoballs (*200*). Single nanoballs are incorporated in each active site of the sequencing flow cell and sequencing is performed either using combinatorial probe-anchor ligation or combinatorial probe-anchor synthesis (*201*).

In 2007 Applied Biosystems commercialised the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) system, which was discontinued in 2016. SOLiD was a sequencing by ligation approach, which utilised emulsion PCR and bead enrichment to produce clonal bead populations of target DNA sequences. The clonal beads generated were deposited onto a glass slide and primer hybridization to the sequencing adapter initiated multiple cycles of ligation, detection and cleavage.

A few years later in 2009, Helicos Biosciences released the first commercial sequencer using single molecule fluorescent technology (HeliScope). This single molecule technology was the first that did not require clonal DNA amplification, a feature shared by all its predecessors (*166*). DNA templates are attached to a surface prior to sequencing by synthesis, similarly to Illumina sequencing. Briefly, fluorescent reversible terminator nucleotides are added, washed and imaged prior to cleavage and repetition of the process with the next nucleotide. Despite the company going bankrupt in 2012 the technology was revived in 2015 with the launch of the GenoCare analyser for clinical applications from Direct Genomics (*202*).

1.4.3 Sequencing developments of the last 20 years

In 2010, the Ion Torrent PGM was released, which performed sequencing by synthesis using ion semiconductor sequencing technology. Sequencing adapted DNA fragments are attached to beads, known as Ion Sphere Particles, and clonally amplified by emulsion PCR. The beads are distributed over the sequencing chip within proton-sensing wells and sequencing is conducted with the introduction of each base sequentially. The protons released when a base is incorporated are detected and the signal generated is proportional to the number of bases incorporated.

Pacific BioSciences commercialised the first single molecule real-time technology (SMRT) in 2011, an approach originally described by Levene *et al.* in 2003, further developed by Eid *et al.* and Pacific Biosciences in 2009 (*203*, *204*). The method monitored DNA synthesis of single molecules in real-time with the use of small light detection volumes called zero-mode waveguides. A DNA template-polymerase complex is immobilised at the bottom of the zero-mode waveguide and sequencing is initiated with the addition of phospholinked nucleotides into the chamber. Nucleotides are labelled with different fluorophores and each time a nucleotide is incorporated a light pulse is produced once the phosphate chain is cleaved and the attached fluorophore is released.

A recent major innovation in the field of nucleic acid sequencing was the methodological advancement of nanopore sequencing introduced in 2012 (*205*). Nanopore sequencing was conceptualised in 1989 by David Deamer and first described in 1996 by Kasianowicz *et al.* (*206*). The authors showed that the measurement of transient current blockages caused by the translocation of a single-stranded DNA or RNA molecule can be used to infer its length. Heptameric transmembrane channels of a-hemolysin with a diameter of 2.6 nm, sufficient to accommodate single stranded DNA or RNA, were embedded in a lipid bilayer membrane. In the absence of a nucleic acid molecule, the application of a potential across the membrane resulted in a continuous current, free of transients. However, in the presence of a polynucleotide, the potential applied presented transient current blockages, which were detected and identified to be proportional to the length of the polymer. The authors noted that with further improvements the method could identify the sequence of bases in a polynucleotide if the changes in ionic current reflected the molecular size and chemical properties of each nucleotide. In 1998, the "Characterisation of individual polymer molecules based on monomer-interface" patent (*207*) was filed. The patent described the rapid and reliable characterization of nucleic acid molecules, both in length and sequence, through the use of an ion-conducting pore or channel and a detection mechanism to monitor the changes caused in the conductance of ions across the pore. The ionic current changes occurring when the polymer passes through the aperture of the channel embedded in an artificial membrane reflect the identities of the polymer monomers. For more than a decade the focus of the research groups involved and later of the company founded in 2005 (Oxford Nanopore Technologies), was the development of nanopore sequencing chemistry and technology for the delivery of commercial nanopore sequencers. In 2014, Oxford Nanopore Technologies launched the first nanopore sequencer to be made available to the scientific community, the MinION: a portable,

palm sized device that allows for real-time long-read DNA and RNA nanopore sequencing. Pore-forming proteins (nanopores) are embedded in an electrically resistant polymer membrane located within the MinION flow cell, which is equipped with 512 detection channels each connected to 4 nanopores. A current is applied through each nanopore by introducing a voltage along the membrane. Once a DNA molecule is close to the aperture of a nanopore it is captured and translocated through, creating a characteristic disruption in the current. For the successful capture and translocation of a DNA molecule, the presence of a motor protein is required, which allows for the separation of the two DNA strands and importantly limits the rate at which the DNA strand translocates through the pore. Each group of bases that goes through the nanopore creates a specific current drop/change, which is detected and measured by a sensor. The signal is then passed to the application-specific integrated circuit (ASIC) and finally to the MinKNOW software. The files generated encode the raw signal information, which is then basecalled by a variety of software that has changed over time. In 2017, Oxford Nanopore Technologies announced the release of the GridION X5, which is a desktop system composed of five MinION Flow cells and integrated computing power, scaling up throughput and analysis capabilities. A year later, the PromethION Early Access Programme was initiated and by April 2019, the platform was reported internally to generate more than 7Tb in a single sequencing run.

The increasing demand for sequencing platforms that convey fast, inexpensive and accurate genome information has encouraged the development of sequencing technologies (*197*). An overview of the key sequencing developments mentioned from 1990 up until 2018 can be found in Figure 1.6. The constant development and improvement of sequencing platforms since 1977 has led to a decrease in cost of sequencing per base and the development of benchtop and mobile sequencers that can be set-up in laboratories easily and cost-effectively (*208*).

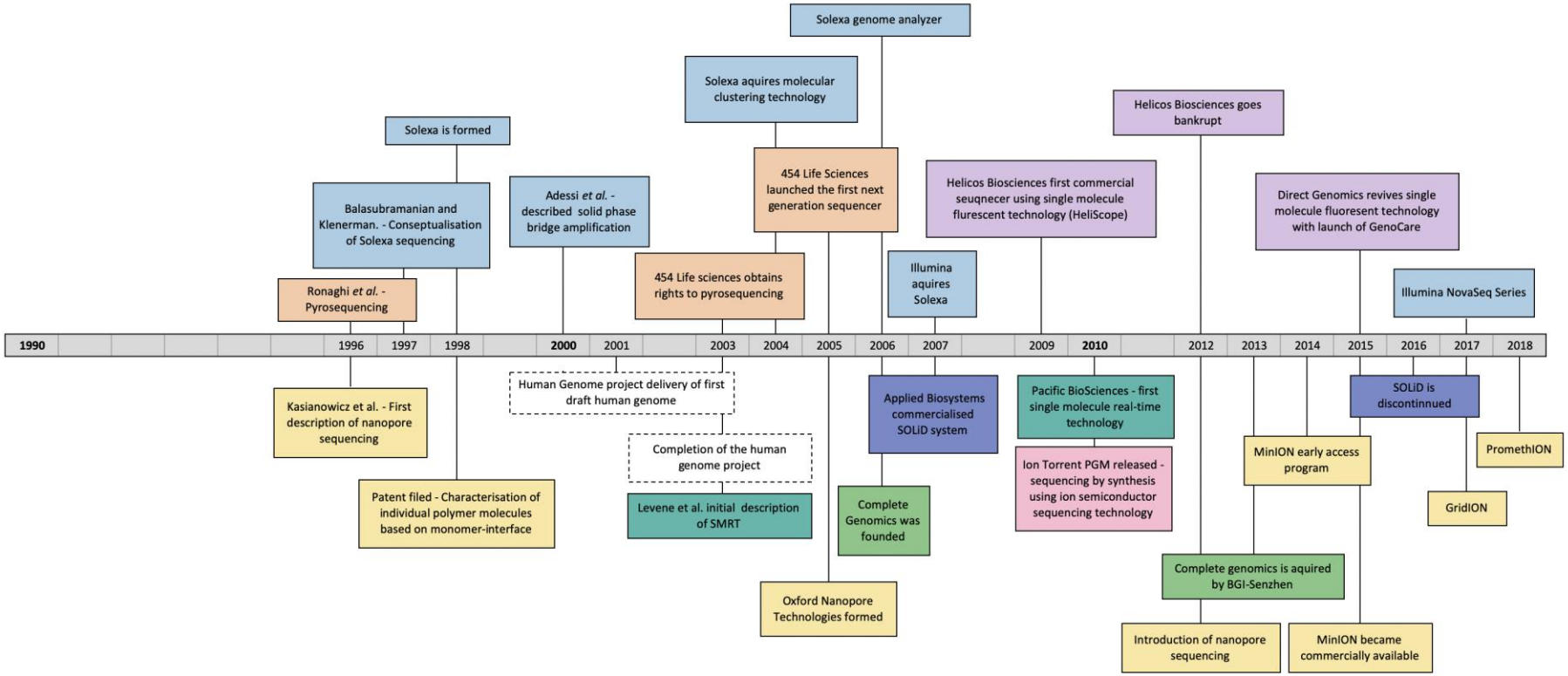


Figure 1.6 Timeline from 1990 up until 2018 of sequencing technology advancements.

A different colour is used for each technology. Illumina Sequencing (Light Blue), Nanopore Sequencing (Yellow), Pyrosequencing (Orange), SOLiD (Dark Blue), Single molecule fluorescent technology sequencing (Purple), SMRT (Light Green), Complete Genomics (Dark Green), Ion Torrent (Pink). Dashed boxes include key events, which occurred as a result of sequencing advancements.

Sanger sequencing was one of the first sequencing methods described and grew to become the most widely used sequencing platform for years, with reads of a few hundred base-pairs initially and later following platform and technological advancements read lengths reaching ~1kb. The increase in data output that came in the decades that followed was majorly driven by short read technologies utilising clonal amplification such as Illumina and Ion Torrent. Ion torrent platforms reach read lengths up to 600 bp and produce data outputs of up to 50 Gb while Illumina platforms generate read lengths of 300bp (paired-end reads) and produce data outputs of up to 6 Tb. Single molecule fluorescent technology introduced by Helicos Biosciences was the first commercial sequencing technology that did not require clonal DNA amplification introducing however the platform produced significantly short read length (median read length ~35 nt). The true utility of single molecule sequencing was realised with the introduction of Pacific BioSciences, which allowed for long read length (~20 kb), high consensus accuracy, capability of epigenetic characterisation and a low degree of sequencing bias. The introduction of nanopore sequencing brought affordable, rapid and accessible single molecule sequencing and an apparent increase in sequencing read length and data output. To date the longest read reported using nanopore sequencing exceeds 2 Mb and the data output of the PromethION was reported to exceed 7 Tb from a single run, comparable in capacity with the equivalent short-read Illumina platform. An overview of the most commonly used platforms and their sequencing specifications can be found in Table 1.1. and an overview of their advantages and disadvantages can be found in Table 1.2.

Table 1.1 Overview of the most commonly used platforms and their sequencing specifications

Technology	Platform	Read length (bp)	Max output per run (reads)	Max output	Run-time
Sanger (*209*)	SeqStudio Genetic Analyzer	800	67 thousand	-	Min 30 min
	3500 Genetic Analyzer	At least 850	138 - 403 thousand	-	Min 30 min
	Refreshed 3730 Genetic Analyzer	900	1.38 - 2.76 million	-	Min 20 min
Illumina (*210*)	iSeq	2 x 150	4 million	1.2 Gb	9.5 - 19 hrs
	MiniSeq	2 x 150	25 million	7.5 Gb	4 - 24 hrs
	MiSeq	2 x 300	26 million	15 Gb	4 - 55 hrs
	NextSeq	2 x 150	400 million	120 Gb	12 - 30 hrs
	NovaSeq	2 x 250	20 billion	6 Tb	13 - 44 hrs
Pacific Biosciences (*211*)	Sequel	Up to 20 kb	500 thousand	Up to 20Gb	Up to 20 hrs
	Sequel II	Up to 20 kb	~ 8 x more data than Sequel		Up to 30 hrs
Ion Torrent (*212*)	GeneStudio	Up to 600	130 million	15 - 50 Gb	6.5 - 19 hrs
	Ion PGM	200 and 400	400 thousand - 5.5 million	60 Mb - 2 Gb	2 - 8 hrs
	Ion Proton	200	60 - 80 million	15 Gb	2.5 hrs
	Genexus	200 - 400	48-60 million	-	-
Nanopore (*213*)	Flongle	Longest read so far > 2Mb	Variable	1 - 2 Gb	1 min - 16 hrs
	MinION			15 - 30 Gb	1 min - 48 hrs
	GrinION			75 - 150 Gb	1 min - 48 hrs
	PromethION			2.4 - 8.6 Tb	1 min - 72 hrs

Table 1.2 Overview of the most commonly used platforms and advantages and disadvantages

Technology	**Sequencing technology**	**Amplification**	**Advantages**	**Disadvantages**
Sanger	Deoxy chain termination sequencing by synthesis	PCR	● Easy of sequencing ● Widely accessible ● Accuracy of sequencing	● Low throughput ● Labour, cost and time intensive for larger studies
Illumina	Reversible terminator sequencing by synthesis	Solid-phase PCR	● High throughput ● Low read error rate ● Low cost per base	● Long run time ● High platform cost ● Short read length ● Large instrument size
Pacific Biosciences	Single molecule, real-time sequencing by synthesis	NA (Single molecule)	● Long read length ● No amplification required ● Low degree of bias ● High consensus accuracy ● Identification of modified bases	● High set-up and platform cost ● Low throughput ● Large instrument size ● Higher read error rate
Ion Torrent	Sequencing by synthesis	Emulsion PCR	● Low platform and running cost ● Short run time	● Requires separate amplification of sequence libraries ● Error rate related to homopolymers ● Poor coverage at AT-rich regions or G-C rich regions
Nanopore	Nanopore	NA (Single molecule)	● Portable ● No platform cost or maintenance ● Long-reads ● High accuracy consensus ● No amplification required ● Real-time data generation	● Currently under continuous development, ● Read error rate (~ 5% with R9.4 flowcells)

1.5. MinION viral sequencing

Virus diagnosis and surveillance has shifted towards rapid generation of complete viral genomes, which can provide clinically relevant information around pathogen identity, drug resistance genotypes and the presence of antigenic determinants. Similarly, the study of viral evolution and molecular epidemiology can be critical for informing public health interventions and infection control (*214*).

Advancements in high throughput sequencing have created new opportunities for viral epidemiology; they allow for identification and characterisation of newly emerging viruses along with detection and monitoring of ongoing viral outbreaks (*100*, *101*). With the incidence of emerging viruses rising in remote and developing countries, the need for rapid diagnostic support in these settings is imperative, however conventional sequencing technologies are limited by their need for complex maintenance, continuous power supply and large equipment footprint. The MinION allows for long-read portable real time DNA and RNA sequencing and overcomes many limitations of conventional sequencing technologies, offering the prospect of combining diagnostics with complete viral genome generation, in real-time, in resource-limited settings. Direct nucleic acid sequencing of clinical samples requires detection of sufficient viral reads to assemble the genome confidently, which can be challenging due to the presence of high abundance host genetic material (*102*). Three main approaches for complete viral genome sequencing from clinical samples have been coupled with portable nanopore sequencing: (i) Target capture enrichment sequencing, (ii) PCR amplification sequencing and (iii) Metagenomic sequencing.

1.5.1 Target capture enrichment sequencing

Target capture enrichment protocols utilise virus-specific capture oligonucleotides, which allow for the enrichment of viral genomes present in a sample prior to sequencing. Target enrichment sequencing on conventional platforms has proven useful for sequencing of emerging viral pathogens in clinical samples (*111*, *215*) and a target enrichment protocol for selective enrichment of large fragments has been demonstrated successfully on the MinION for cultured influenza A virus (*216*) and DENV (*217*). These methods offer low limits of detection, but can suffer from coverage bias, are more expensive and require lengthy protocols (typically a 12-24hr hybridisation step) (*216*, *218*). It is an approach that requires advanced design and *a priori* knowledge of the target pathogen, and is more costly and time-consuming

compared to PCR amplicon and metagenomic sequencing (*102*). Target enrichment can be extremely useful in a clinical setting but is challenging to establish for rapid sequencing during a virus epidemic or outbreak in less-equipped settings.

1.5.2 PCR amplification sequencing

PCR amplicon protocols allow for selective amplification of a viral genome of interest within a clinical sample prior to sequencing. PCR amplicon enrichment generates amplified viral genome fragments utilising primers complementary to a previously known viral nucleotide sequence (*102*). Sequencing PCR amplicons is a sensitive, cost efficient approach and is particularly useful for samples with low viral load, however, it is a laborious approach to scale up for a large amount of samples when a high number of amplicons are required per sample (*156*, *218*). Its application in recovery of whole genome sequences of an emerging virus was highlighted in its use during the West Africa EBOV 2013-2016 epidemic. Quick *et al.* in 2016 used a tiling PCR approach, applying a combination of 11 (1.3 - 2.4 kb length) or 19 (0.9 - 1.8 kb length) separate amplicon reactions coupled with nanopore sequencing to reliably recover more than 97% of the EBOV genome (18.959 kb) for 142 EBOV RT-PCR positive samples (*147*). Furthermore, amplicon sequencing was used during the ZIKV epidemic in the Americas in 2015 (*156*, *218*).

These efforts resulted in the development of a novel multiplex PCR enrichment protocol which overcomes difficulties encountered with single-plex approaches, mainly the challenge of sequencing a high number of samples that require shorter fragment amplification, which is the case for viral clinical samples with high Ct values (>30), such as ZIKV positive samples (*156*, *218*). The protocol described is an easily scalable approach, which allows for amplification of the whole virus genome using a multiplexed tilling PCR implemented in only two reactions; reducing reagent costs, manual pipetting steps, hands-on time and minimizing laboratory contamination and errors. PCR amplicon sequencing approaches have also been successfully used for YFV in Brazil (*219*) and WNV in the USA (*219*, *220*). However, targeted methods are prone to amplicon contamination and are limited in coverage recovery of lower abundance regions and those regions which fall outside the primer pair coverage (both the 3' and 5' UTR regions) (*218*). The design of PCR amplicons offers many advantages but the need for previous knowledge of the sequence of the virus of interest leads to great challenges when encountered with highly diverse or recombinant viruses and specificity for the virus it has been designed against does not allow for viral discovery or unguided identification (*102*, *218*).

1.5.3 Metagenomic sequencing

Metagenomics was originally defined as the analysis of an environmental sample to determine the sequences of the collective microbial genomes contained in it (*221*). The use of the term has since been extended to refer to any culture-independent analysis of microbial communities following the application of sequencing approaches in a variety of fields, including viral genomics (*222*). Clinical metagenomic protocols permit untargeted sequencing of total nucleic acid within a sample, including viral, host, bacterial and fungal genomic material that could be present (*223*). This approach is especially promising in clinical diagnosis, viral discovery and identification of viral co-infections (*224*). Diagnostic viral metagenomics has identified the causative agent of outbreaks and led to the discovery of novel viruses via a variety of metagenomic protocols (*225*–*229*). The success of a metagenomic approach is highly dependent on the method's sensitivity when performed on a background-rich sample, both to detect and also to fully characterise the genome of the pathogen identified (*230*). This is one of the main limitations of the approach as it generally requires high sequencing depth to ensure sufficient viral reads are obtained, however, with advances in sequencing technologies and the decrease in sequencing cost the potential for pathogen discovery using metagenomic sequencing has greatly improved (*231*).

1.6 Sequence-independent single primer amplification

Metagenomic protocols require no *a priori* knowledge of the pathogen of interest for their design, offering a hypothesis-free approach to target identification. Sequence-independent single primer amplification (SISPA) methods are commonly used metagenomic protocols for viral detection and discovery and can be grouped in two categories (a) SISPA by ligation (b) SISPA by random PCR.

1.6.1 SISPA by ligation

The ligation mediated SISPA method was originally developed by Reyes and Kim in 1991 for the amplification of heterogeneous DNA populations (*232*). Extracted RNA is reverse transcribed and converted to double stranded cDNA with the use of random or oligo-dT priming of the first strand. The method subsequently entails the directional ligation of an asymmetric adapter, called a linker/primer, onto the blunt ends of the resulting cDNA population using T4 DNA ligase. The primer sequence present on each end of the ligated DNA is used to enrich the cDNA population by PCR. The resulting amplified material can be used for downstream processing. In the original description of the ligation mediated SISPA method, restriction enzyme sites (EcoRI, NotI) were incorporated within the linker/primer. The resulting amplified material from the SISPA step could be incorporated into vectors by molecular cloning utilising the restriction endonuclease sites and sequenced. An overview of the SISPA by ligation method can be found in Figure 1.7. The protocol was initially evaluated on bacteriophage phi X174 and it was demonstrated that amplification was successful even at the lowest input amount tested. The method was used for cDNA preparation from low abundance sequences of Hepatitis C virus present in a positive serum sample. In 1991, Reyes *et al. (233)*, used the SISPA method for the identification of the cause of a non-A and non-B hepatitis epidemic outbreak in humans, a new virus species: Hepatitis E virus. Later, Matsui *et al. (234)* coupled the SISPA method with conventional cloning and immunoscreening techniques for the identification of cDNA clones originating from a protein coding region of the Norwalk virus genome. Molecular characterisation of the virus had previously been difficult due to its low abundance in patient samples and its detection was limited to immunological assays. The use of non-selective cDNA amplification led to the identification of a portion of the viral genome and created opportunities for better characterisation of the virus. A variation of the method was introduced in 1992 by Lambden *et al. (235)* for dsRNA viruses. The linker/primer is directly ligated to the dsRNA, the dsRNA strands are

separated and reverse transcribed into cDNA. The resulting double stranded cDNA is amplified with the SISPA step in the same fashion as the original protocol.

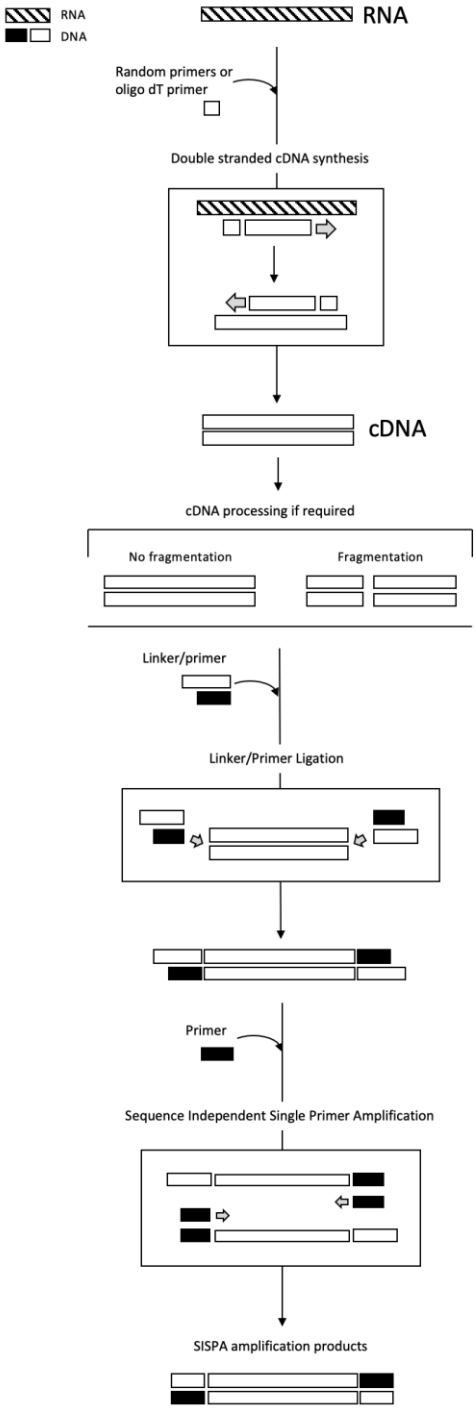


Figure 1.7 Overview of SISPA by ligation method.

As originally described (232, 236), extracted RNA is reverse transcribed and converted to double stranded cDNA with the use of random or oligo-dT priming of the first strand. The method subsequently entails the directional ligation of an asymmetric adapter, called a linker/primer, onto the blunt ends of the resulting cDNA population using T4 DNA ligase. The

primer sequence present on each end of the ligated DNA is used to enrich the cDNA population by PCR.

1.6.2 SISPA by random PCR

SISPA by random PCR involves the use of a tagged random primer for the amplification of a cDNA population. The approach was first described in 1992 by Froussard (*237*), who tested it on MS2 phage RNA. In the first step of the protocol, RNA is reverse transcribed using a 26 nucleotide primer containing a random hexamer at its 3' end. The random hexamer leads to a mixture of 6^4 (4096) different primer combinations which anneals to different locations along the RNA sequence and a complementary DNA strand is synthesised. Following the first strand cDNA synthesis, DNA polymerase I Klenow fragment is added to the reaction for the second strand cDNA synthesis, also primed by the 26 nucleotide primer containing a random hexamer at its 3' end. The dsDNA generated therefore contains the tag sequence at both ends. SISPA of the randomly synthesised double stranded cDNA is performed in the presence of a single primer containing only the tag sequence. An overview of the SISPA by random PCR method can be found in Figure 1.8. Following SISPA the random amplified sequences could be visualised and size selected on an agarose gel, if a specific length range was desired. The resulting fragments, whether size selected or not, were purified, cloned and sequenced. Bohlander *et al.* (*238*) within the same year described the application of SISPA by PCR for direct amplification of microdissected chromosomal material. The approach included two rounds of DNA synthesis using DNA polymerase and a 21 nucleotide primer containing a random pentamer at its 3' end. Flanking of the target DNA population with tagged random primers allows for SISPA amplification of unknown sequences with the use of the known sequence (tag-primer) incorporated. The amplified product was biotinylated and used for fluorescence in situ hybridisation to confirm their chromosomal location.

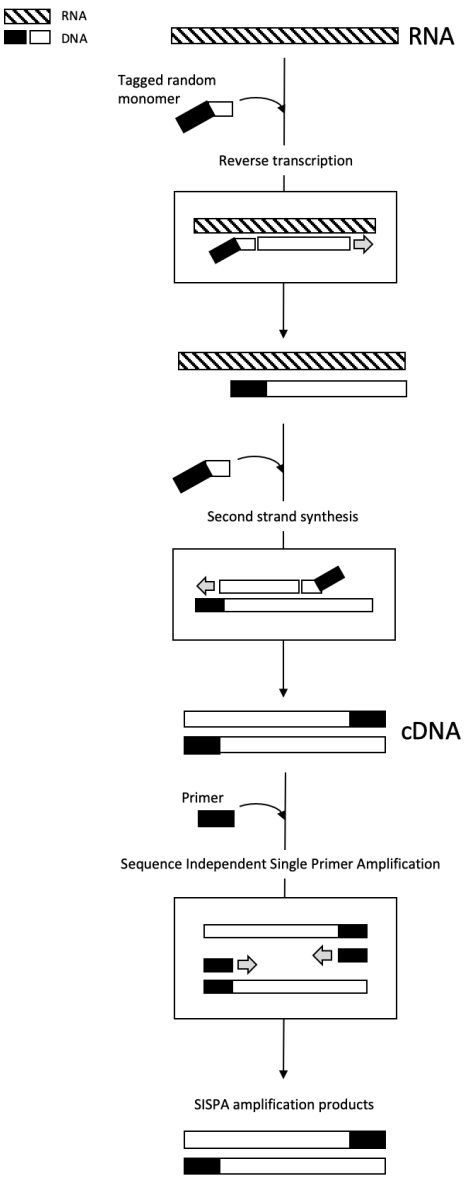


Figure 1.8 Overview of SISPA by random PCR method.

As originally described (236, 237), the random primer anneals to different locations along the RNA sequence and a complementary DNA strand is synthesised. Following the first strand cDNA synthesis, DNA polymerase I Klenow fragment is added to the reaction for the second strand cDNA synthesis, also primed by the 26 nucleotide primer containing a random hexamer at its 3' end. The dsDNA generated therefore contains the tag sequence at both ends. SISPA of the randomly synthesised double stranded cDNA is performed in the presence of a single primer containing only the tag sequence.

1.6.3 Recent applications for human viral pathogens

Over the past two decades, the SISPA protocol has been widely used for viral metagenomic sequencing. Different 5' tag-primer sequences and various lengths of 3' random sequence have been used. The tag-sequence commonly still contains restriction enzyme sites, which may facilitate downstream molecular cloning and sequencing, however recently the majority of methods proceed directly to sequencing following the SISPA.

Wang *et al.* (*239*) in 2002 used a SISPA strategy in conjunction with microarray analysis for the detection of human viruses in respiratory samples. The approach utilised a tagged random octamer for the reverse transcription of the extracted RNA and Sequenase enzyme for the second strand DNA synthesis. The double stranded cDNA generated was amplified using a primer targeting the tag sequence and the products hybridized onto a custom designed DNA microarray. The DNA microarray included oligonucleotides representing more than 140 viruses, containing all available respiratory tract virus sequences available at the time, and was validated against respiratory virus culture extracts. Subsequent analysis of nine patient samples with respiratory infection symptoms led to clear indication of rhinovirus infection in four samples and of parainfluenza 1 in one sample, the presence of both viruses was confirmed by PCR. Ksiazek et al. (*132*) used an improved version of the DNA microarray coupled with the SISPA by PCR protocol during the 2003 outbreak of SARS, later described in detail by Wang et al. (*240*). The protocol was successfully used to identify the presence of a previously uncharacterized CoV in a viral culture isolate from a SARS patient. At the same time, Marra et al. (*241*) sequenced the complete genome of the SARS–associated CoV using a ligation-mediated SISPA approach coupled with molecular cloning and Sanger sequencing. In 2007, a novel Polyomavirus was identified from acute respiratory tract infection samples, using a SISPA by PCR protocol (*242*). Extracted RNA was reverse transcribed in the presence of a tagged random nonamer and cDNA was amplified by SISPA using the tag sequence. A year later, Towner *et al. (243)* used a SISPA by PCR protocol as described by Cox-Foster *et al. (244)* to investigate the cause of a haemorrhagic fever outbreak. The random-primed protocol approach was coupled with pyrosequencing to investigate reported cases of haemorrhagic fever in Western Uganda. Sequencing of total RNA extracted from a patient serum sample led to the identification of a novel ebolavirus species (Bundibugyo) and the recovery of approximately 70% of the virus genome. Similarly, a novel haemorrhagic fever–associated arenavirus, Lujo virus, was identified in 2008 using the same

unbiased high-throughput pyrosequencing approach (*245*–*247*). In 2009 Greninger et al. (*248*) demonstrated the utility of the SISPA by PCR protocol coupled with two different strategies, a pan-viral microarray assay and deep sequencing, for the identification and characterisation of H1N1 influenza A virus. The microarray approach successfully differentiated the seasonal H3N2 and H1N1 influenza from the newly emergent 2009 H1N1 pandemic strain. Deep sequencing of 17 clinical samples collected during the 2009 H1N1 pandemic enabled the detection of the virus in all samples and reached 97% virus genome coverage for one of the samples. Chen et al. (*101*) in 2011 described the detailed protocol involving SISPA by PCR prior to hybridization to the pan-viral microarray (Virochip) assay for clinical sample screening. The Virochip at the time consisted of approximately 36,000 probes derived from over 1,500 viruses available on Genbank. A SISPA by PCR protocol using a tagged random octamer (*249*) coupled with deep sequencing was for the retrospective investigation of three acute haemorrhagic fever cases presented in 2009 in the Democratic Republic of Congo. The investigation led to the identification of a novel haemorrhagic fever associated Rhabdovirus, Bas-Congo virus, in 2012 (*250*). Metagenomic sequencing using the SISPA by PCR protocol coupled with the Oxford Nanopore MinION for viral pathogens, was successfully demonstrated in principle in 2015. Greninger *et al.* (*251*) reported the unbiased detection of CHIKV, EBOV, and hepatitis C from human blood samples. Extracted RNA from all samples was reverse transcribed using a tagged random nonamer followed by second-strand synthesis and SISPA. The amplified cDNA was subjected to nanopore sequencing using the MinION. Within ~2hr of the sequencing run onset, a total of 6.7% of reads mapping to CHIKV and 90% of the genome was recovered. EBOV and hepatitis C samples both generated a smaller amount of reads mapping to the respective virus but nevertheless permitted the identification of the correct viral strain.

The potential of metagenomic protocols has been evident for several years. Advancements in sequencing technology, such as increased data output and reduced and sample-to-results turnaround times make the coupling of metagenomic protocols with NGS more accessible and feasible in a variety of clinical environments.

1.7 Lassa virus: A closer look

1.7.1 Background

Lassa fever was first described in 1969 in the town of Lassa, Nigeria (*40*). A missionary nurse working in the village fell ill from a previously unknown haemorrhagic fever (*252*). She was transported to Bingham Hospital, Jos, Nigeria where a Bingham missionary nurse became infected as a result of human-to-human transmission whilst caring for her. Both missionary nurses became fatally ill. Both had been taken care of by a third missionary nurse, Lily Pinneo, who was also infected. Due to the fact that she was the third case to present similar symptoms, she was transported back to New York for diagnosis and treatment (*40*, *252*). She eventually made a full recovery making her the first known patient to survive the new haemorrhagic fever disease. Blood samples taken from all three cases were sent to Yale Arbovirus Research Unit, Yale University to virologist Jordi Casals, who was the first to isolate the virus (*253*). The Yale Arbovirus Research Unit laboratory reported that all initial serum samples had a cytopathic effect on tissue culture and plaques were observed within 5-8 days. Electron microscopy identified spherical particles resembling those previously reported for lymphocytic choriomeningitis virus, a virus of the same genus (*Mammarenavirus*). Young adult mice were found to become sick after inoculation, similar to lymphocytic choriomeningitis virus (*254*). The virus was calculated by filtration to have a diameter of 70-150 nm and the viral genome appeared to be RNA. Casal and a laboratory technician, Juan Roman, both became infected, the latter fatally. Casal was treated with Lily Pinneo's convalescent serum and recovered. The novel arenavirus was isolated from the blood samples of the missionary nurses and that of Casal. The disease (Lassa fever) and virus (Lassa mammarenavirus) were assigned their name after the Nigerian town, Lassa, the location where the index case worked and most likely contracted the viral infection (*40*, *255*).

1.7.1.1 Symptoms

Symptoms of Lassa fever are typically presented within 7-21 days of infection with LASV (*55*). The high population seroprevalence of LASV-specific antibodies in LASV endemic areas indicates that the majority (~80%) of infections are mild or possibly asymptomatic and in most cases do not lead to patient hospitalisation (*55*). However the acute viral haemorrhagic illness is estimated to affect between 300,000

and 500,000 people annually (*256*, *257*) with a typical CFR of 18%, although significant increases of up to 31% have been reported during endemic seasons or among hospitalised patients (*39*).

The onset of the disease is gradual with the early stages resembling other common diseases such as influenza or malaria (*258*). When symptomatic, the initial clinical presentation is characterised by fever, weakness and malaise. After a few days additional symptoms such as cough, red eyes, sore throat, severe headache, chest pain, muscle pain, abdominal pain, nausea, vomiting, diarrhoea may occur (*259*). Severe forms of the disease are commonly associated with multi-organ complications such as facial edema, fluid in the lungs, low blood pressure and bleeding from the mouth, nose, vagina or gastrointestinal tract (260). Additional symptoms that have been documented at later stages of the disease are shock, seizures, tremor, disorientation, coma and deafness. Deafness has been identified to occur in ~25% of patients that survive and in half of the cases, partial hearing recovery occurs after 1-3 months. In fatal cases, death commonly occurs within 14 days of onset. The disease is particularly severe during the third trimester of pregnancy, with more than 80% of cases leading to a fatal outcome for the mother and/or the foetus (*261*).

1.7.1.2 Diagnosis

Lassa fever diagnosis in endemic areas or in patients returning from endemic areas is based on clinical features and/or laboratory confirmation. Standard case definitions have been defined by the WHO/Regional Office for Africa (WHO/AFRO) and the CDC in the "Technical Guidelines for Integrated Disease Surveillance and Response in the African Region (IDSR)" document (*262*). It states that Lassa fever suspect cases are those, which present disease with a gradual onset of one or more of the following symptoms: malaise, fever, headache, sore throat, cough, nausea, vomiting, diarrhoea, myalgia, chest pain hearing loss combined with a history of contact with rodents, their excreta or with a known human case of Lassa fever. Confirmed cases are defined as those, which satisfy the suspect case definition and have been laboratory confirmed (positive IgM antibody, PCR or virus isolation) or epidemiologically linked to a Lassa fever confirmed case.

Clinical diagnosis is often hard due to the similarity of Lassa fever symptoms to those of other common diseases (*263*). This can often lead to misdiagnosis, especially early in the course of the disease. Additionally Lassa fever can be difficult to distinguish from other haemorrhagic fevers or diseases that cause fever such as

EBOV disease, malaria and typhoid fever (*264*, *265*). Definitive diagnosis requires laboratory diagnosis, which can be achieved by a range of techniques such as RT-PCR, antibody enzyme-linked immunosorbent assay (ELISA) and antigen detection. The current method of choice for the early detection of LASV in clinical samples is RT-PCR (*39*, *266*, *267*). The first RT-PCR assays used to detect LASV were developed using a limited number of genome sequences that were available at the time (*268*, *269*). The technological advances in PCR and the increase in number of sequenced LASV samples has had a dramatic impact on diagnostic capabilities, however the diversity of the virus still continues to pose challenges for its detection (*50*, *51*). The assay currently considered the gold standard is the commercially produced RT-PCR assay by Altona RealStar®, the first version of the assay targeted the S segment only, but the recently released updated version includes two assays, one targeting the L and the other the S segment (*270*).

1.7.2 Zoonotic reservoir and transmission

The primary animal reservoir of LASV is the multimammate rodent, *Mastomys natalensis (M. natalensis)*, commonly known as the "multimammate rat", due to the female's multiple and prominent mammary glands. *M. natalensis is a* dominant species in the ecosystem of Sub-Saharan Africa distributed across a variety of habitats, including savannahs, woodlands, forested areas and, importantly, houses and cultivated fields (*271*, *272*). This varied habitat distribution demonstrates its capability to quickly adapt (*273*). Once infected, the rodents present persistent, asymptomatic infection and shed virus through their urine, faeces and saliva (*274*). Despite the presence of the *Mastomys* rodents across Sub-Saharan Africa, Lassa fever has not been identified outside West Africa (*275*).

LASV was first identified in *M. natalensis* in 1972 and its role as a reservoir host confirmed by subsequent studies in Nigeria in 1975 (*276*). A study in 1983 investigated the presence of *M. natalensis* rodents in Sierra Leone households where Lassa fever cases occurred (*277*). They reported that 79% of all rodents caught within the houses were *M. natalensis* and 39% of those were viremic compared to 3.7% in control houses. In 1987, McCormick *et al. (278)* surveyed populations in 15 different Sierra Leonean villages and found that human seroprevalence of anti-LASV antibodies ranged from 8% to 52%. They identified a high prevalence of infestation with *M. natalensis* in villages and that the rodents were found more frequently in houses than the surrounding areas. This finding suggested that humans are likely infected in their homes through primary exposure to the rodents or their excreta. In

endemic regions 50-100% of rodents caught in houses have been identified to the genus *Mastomys,* with a large fraction carrying LASV (*277–280*). Lecompte *et al.* (2006), screened 1,482 murid rodents caught across Guinea and identified LASV infections only in *M. natalensis* species (*80*). Additionally, *Mastomys* rodents are known to be a food source, thus infection may occur during catching or meal preparation. Recently additional rodent reservoirs of LASV have been identified, *Mastomys erythroleucus (278, 281)* and *Hylomyscus pamfi (281)* However, the role of *M. natalensis as* the primary reservoir continues to be strongly supported by the literature.

Human to human LASV transmission events have also been reported, through contact with bodily fluids of an infected patient, primarily as nosocomial infections and less often as community infections (*46*, *47*, *282*). Nosocomial transmission occurs when health workers come into contact with infected blood or bodily fluids, with surgical procedures and births both particularly high risk sources of infection (*283*, *284*). Community transmission events can occur through exposure to infected blood or bodily fluids; this is commonly family members taking care of a Lassa fever positive relative (*275*). Secondary transmission during patient convalescence has not been documented, despite the delayed clearance of the virus. The only exception being rare reports of sexual transmission occurring months after patient recovery from acute disease (*275*, *285*). Extensive human-to-human transmission has not been identified and its incidence is extremely low.

In 2016, Fichet-Calvet *et al. (286)* highlighted the importance of rodent control efforts to cover large geographical areas and the need to sustain them over time. Reduction in Lassa fever cases is currently primarily driven by rodent control, community hygiene and human awareness and behaviour (*287*).

1.7.3 Molecular surveillance and epidemiology

Between 1969 (first case) and 1984, Lassa fever grew as a major public health concern after five epidemics in four West African countries: Nigeria (*40*, *288*, *289*), Liberia (*290*), Guinea (*291*) and Sierra Leone (*292*). The virus has now been identified in Benin, Togo, Côte d'Ivoire and Mali (*39–44*, *293–295*). The virus is endemic in Nigeria, Liberia, Guinea, Sierra Leone, Mali and Côte d'Ivoire. Individual cases have been identified in Benin and Togo but additional evidence is required to confirm endemicity of the virus (*296*).

LASV is genetically diverse with several different lineages in circulation in West Africa. The four first lineages were described in 2000 by Bowen *et al. (41)* using

a dataset of 49 unique partial nucleoprotein gene sequences. Sampling included samples from Guinea (human (n=4), rodent (n=1)), Liberia (human (n=9)), Nigeria (human (n=15)) and Sierra Leone (human (n=14), rodent (n=11)) taken between 1981 and 1996. Three lineages were identified circulating in Nigeria. Lineage I includes the prototype Pinneo strain, isolated in 1970. Lineage II contained sequences from southern central Nigeria and lineage III, which contained sequences from sites in northern central Nigeria. The fourth lineage contained the largest group of LASV sequences used in the study, all of which originated from Guinea, Liberia and Sierra Leone. The authors noted that the interlineage genetic distances among the three Nigerian lineages (I, II, III) is greater than the distance between lineage III and lineage IV found in Guinea, Liberia and Sierra Leone. Overall variation in the partial nucleoprotein gene sequences used was found as high as 27% at the nucleotide level and up to 15% at the amino acid level. The genetic distance between LASV strains was strongly correlated with geographical distance suggesting that the reservoir (*Mastomys sp.*) had exhibited minimal regional movement since the first identification of the virus in 1970. Finally, the authors did not find evidence of a molecular clock, evolution occurring in a detectable constant rate. However, they highlight that it is likely a molecular clock is present in LASV evolution but was not evident due to the short time frame of data sampling (1969-1997).

In 2000, Gunther *et al.* (*44*) isolated and characterised a novel LASV strain, imported into Germany by a traveller who had visited Ghana, Côte D'Ivoire, and Burkina Faso. The patient had visited all three countries during the 21 day incubation period, thus the exact origin of the sequence could not be identified. Evidence of LASV in this region of West Africa only existed historically from Burkina Faso, from a single imported non-fatal case in the Netherlands in 1980 (*297*). In 2009, Atkin *et al.* (*43*) reported the first case of LASV from Mali. The patient was medically evacuated from Mali to London for treatment but rapidly deteriorated with a fatal outcome. Phylogenetic analysis revealed that the Mali sequence was distinct from other Lassa strains but clustered most closely with the strain from the imported German case in 2000 (*43*, *44*). In 2010, Ehichioya et al (*298*) investigated the molecular epidemiology of LASV in Nigeria. Phylogenetic analysis of complete GP, NP and L gene sequences confirmed the existence of the previously described four LASV lineages, highlighted the predominance of lineage II and III sequences and postulated the existence of a previously undescribed lineage.

In 2010, the presence of LASV infected rodents and epizootic viral transmission was confirmed in the village of Soromba, Côte D'Ivoire by Safronetz et al (*294*). The sequence obtained from the rodent Soromba LASV was

indistinguishable from the imported Mali case in the UK. In 2011 the first cases of LASV originating in Ghana were reported, the three geographical distinct cases had no travel history outside the area suggesting the infections were all acquired locally (*299*). In 2013, Kouadio *et al.* (*295*), identified LASV positive *Mastomys* rodents in Côte D'Ivoire and subsequently in 2015, Mateo *et al. (300)* reported a LASV fatal human case in western Côte D'Ivoire. Ecologically, southern Mali, southern Burkina Faso, northern Côte D'Ivoire, Ghana, Togo and Benin share common characteristics with the known LASV endemic regions of Nigeria, Guinea, Sierra Leone and Liberia (*301*). All of the aforementioned countries and regions are part of the Tropical Wooded Savanna ecological zone and the environmental factors within these areas might favour the presence of *M. natalensis* and the circulation and spread of LASV (*42*, *301*). In 2015, Andersen *et al.* sequenced ~200 LASV positive samples from rodents and humans from across West Africa. The existence of four LASV lineages was confirmed and samples from rodents and humans were found to cluster together, with no evidence for host-specific clades. High levels of nucleotide diversity were identified, up to 32% for at strain level and 25% for across individual segments. Evidence of segment reassortment was identified but no recombination events were found. Molecular dating identified that LASV most likely originated in Nigeria and subsequently spread into neighbouring countries. A molecular clock was estimated from plotting root-to-tip distances against sample collection dates across the entire LASV dataset. This allowed for the estimation of the time to the most recent common ancestor to be ~1000 years for the L segment and 650 years for the S segment. Whilst evidence of rodent-to-rodent transmission was identified by clustering of the sequences, only 5 of 169 LASV sequences originating from humans resulted in such clusters, leading to the conclusion that most human LASV infections are independent transmissions from the rodent reservoir rather than due to human-to-human transmission. Due to the high diversity identified within different LASV lineages, continuous surveillance and sampling of LASV positive rodents and humans is critical for the development of diagnostic tools, therapeutics and vaccines and for our understanding of LASV evolution and molecular epidemiology.

Since the original four lineages described by Bowen *et al (41)*, LASV lineages have been expanded to include an additional three as proposed currently in the literature. Strains from Mali and Côte D'Ivoire have been proposed to represent lineage V (*302*). A LASV from Nigeria found in a *Hylomyscus pamfi* rodent has been proposed to belong to a new lineage VI (*281*) and a LASV strain from a nosocomial outbreak in Togo has been proposed as lineage VII (*293*). An overview of the current LASV lineages and the region they have been identified in can be found in Figure 1.9.

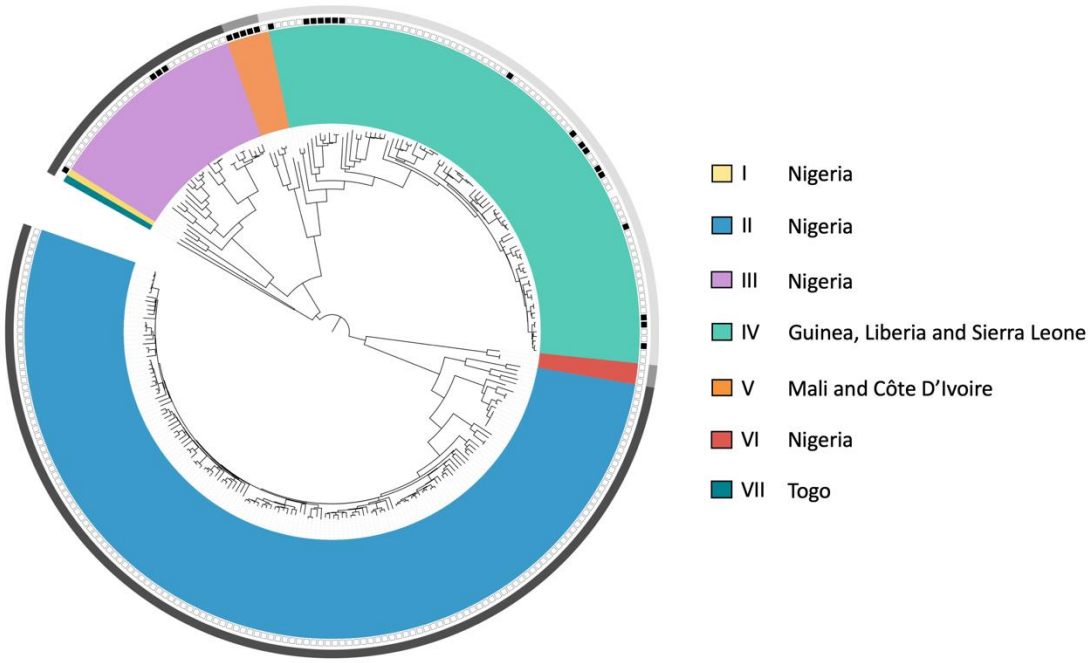


Figure 1.9 Graphical representation of known LASV lineages.

The figure is adapted from the phylogenetic tree of the LASV L segment as reported in (303). The seven genotypes are indicated with different colours and roman numerals. The locations where each lineage has been identified is also included. The colour strip highlights the human LASV sequences obtained from previous years (light grey); sequences obtained from rodent samples (dark grey). The squares at the tips represent the host: empty square = human virus; black square = rodent virus; no square = laboratory virus.

1.8 Thesis scope

This PhD research project set-out to evaluate metagenomic sequencing for the identification and complete genome recovery of pathogenic RNA viruses directly from clinical samples; to interrogate their feasibility and sensitivity when coupled with the Oxford Nanopore MinION and to evaluate its application in resource-limited settings for real-time genomic surveillance and molecular epidemiology of outbreaks.

Chapter 2

Materials and methods

2. Chapter 2. Material and methods

All chemicals were from Sigma unless stated otherwise, Diethyl pyrocarbonate (DEPC) treated nuclease free water (NFW) was from Thermo Fisher Scientific and oligos were from Integrated DNA Technologies (IDT) unless stated otherwise.

2.1 Virus sample collection

HAZV stock (Catalogue No. 0408084v) was obtained from the National Collection of Pathogenic Viruses (Public Health England, UK). Mock-clinical samples were produced by dilution of HAZV in human serum (S7023, Sigma-Aldrich). CHIKV and DENV positive clinical samples were obtained from the Rare and Important Pathogens Laboratory (RIPL, Public Health England, UK). Samples from RIPL tested in this study were residual diagnostic samples used for assay and method development thus required no further ethics approval. LASV samples were acquired through collaboration with the Bernhard Nocht Institute for Tropical Medicine (BNITM), Hamburg, Germany and originated from the Irrua Specialist Teaching Hospital (ISTH), Irrua, Nigeria. All diagnostic samples tested positive by real-time RT-PCR for the virus of interest and were anonymised prior to any investigation. The use of LASV diagnostic leftover specimen and corresponding patient data was approved by the ISTH Research and Ethics Committee (approval ISTH/HREC/20171208/45)

2.2 DNA quantification

The Qubit® Fluorometer (Thermo Fisher Scientific) and dsDNA High Sensitivity (HS) assay kit (Q32854) were used, which has a quantitation range of 0.2-100 ng. The Qubit® working solution is prepared by diluting the Qubit® dsDNA BR Reagent 1:200 in Qubit® dsDNA BR Buffer. For each standard, 190 µL of Qubit® working solution is mixed with 10 µL of Qubit® standard and for each sample, 199 µL of Qubit® working solution is mixed with 1 µL of sample. All tubes are incubated for 2 minutes prior to measurement. Qubit™ Assay Tubes (Q32856, Thermo Fisher Scientific) were used for all Qubit® assay readings and the instrument was calibrated using the standards prior to every use.

2.3 Virus detection and quantification

All RIPL assays used were previously developed and clinically validated in certified diagnostic laboratories for ISO 15189 compliance, with the exception of the LASV Altona assay, which is commercially available (642013, Altona). Details on primer information along with the target and kit used can be found in Table 2.1. All assays were run on the LightCycler 480 (Roche Life Science), on LightCycler 480 optical white reaction plates and sealing foil (Roche, UK). Bacteriophage MS2 was included as a positive control on each run for each and sterile water was used as a negative control.

Table 2.1 Overview of oligonucleotide sequences, kit information, target region and corresponding reference for each assay used

Virus	Oligo	Sequence (5'- 3')	Kit, Manufacturer	Positive control	Target	Ref[a]
HAZV	Forward Primer	CAA GGC AAG CAT TGC ACA AC	QuantiFast SYBR green PCR kit, Qiagen	Cell culture grown virus	S Segment	(*304*)
	Reverse Primer	GCT TTC TCT CAC CCC TTT TAG GA				
CHIKV	Forward Primer	TCG ACG CGC CCT CTT TAA	Superscript III Platinum One-Step Quantitative RT-PCR kit,Invitrogen	synthetic plasmid control	E1 gene	(*305*)
	Reverse Primer	ATC GAA TGC ACC GCA CAC T				
	Probe	6FAM ACC AGC CTG CAC CCA TTC CTC AGA C BHQ1				
DENV	Forward Primer	GGA TAG ACC AGA GAT CCT GCT GT	Superscript III Platinum One-Step Quantitative RT-PCR kit,Invitrogen	synthetic plasmid control	3' non-coding region	(*306*)
	Reverse Primer	CAT TCC ATT TTC TGG CGT TC				
	Reverse Primer 2	CAA TCC ATC TTG CGG CGC TC				
	Probe	6FAM CAG CAT CAT TCC AGG CAC AG BHQ1				
LASV	Altona	Commercial kit (LASV Probe labelled with 6FAM and Internal Control Probe labelled with JOE)	RealStar® Lassa Virus RT-PCR Kit 1.0, Altona	synthetic construct	S Segment	NA
LASV	Forward Primer	CCA CCA TYT TRT GCA TRT GCC A	Superscript III Platinum One-Step Quantitative RT-PCR kit,Invitrogen	synthetic construct	L Segment	*(*307*)
	Reverse Primer	GCA CAT GTN TCH TAY AGY ATG GAY CA				
	Probe	FAM-AAR TGG GGY CCD ATG TGY CCW TT-BBQ				
MS2	Forward Primer	TGG CAC TAC CCC TCT CCG TAT TCA CG	Superscript III Platinum One-Step Quantitative RT-PCR kit,Invitrogen	synthetic plasmid control	Maturation protein coding region	
	Reverse Primer	GTA CGG GCG ACC CCA CGA TGA C				
	Probe	Cy5 CAC ATC GAT AGA TCA AGG TGC CTA CAA GC BBQ1				

[a]Abbreviation for reference and refers to publication associated with the equivalent assay.

2.3.1 HAZV assay

HAZV was quantified using a semi-quantitative two-step RT-PCR (*304*). cDNA was synthesised from 6 µL of RNA using Superscript III First-Strand Synthesis Supermix (Thermo Fisher Scientific) as per manufacturer's instructions and primers listed in Table 2.1. All HAZV PCR assays were run using the QuantiFast SYBR green PCR kit (Qiagen) and PCR conditions, as listed in Table 2.2 and Table 2.3. Standards with a range of 10^7 – 10^1 viral genome copies per ml were included in each run to allow for copy number approximation. Standards were 100 bp amplicon, generated using HAZV specific primers. The PCR product was quantified using the Qubit Fluorometer, the amplicon concentration was calculated and dilutions were made in NFW.

Table 2.2 *HAZV assay reaction mix composition*

Reagent	**Single Reaction**	**Final Concentration**
Reverse Transcription Step 1.		
Primer (Forward)	1 µL	0.2 µM
Annealing Buffer	1 µL	NA
Template	6 µL	NA
Water	2 µL	NA
Reverse Transcription Step 2.		
2X First-Strand Reaction Mix	10 µL	1x
SuperScript™ III/RNaseOUT™ Enzyme Mix	2 µL	NA
PCR		
QuantiFast SYBR green	5 µL	NA
Forward Primer	1 µL	1µM
Reverse Primer	1 µL	1µM
Water	2 µL	NA
Template	1 µL	-
Total	10 µL	-

Table 2.3 HAZV two step RT-PCR cycling parameters

Step Name	Analysis Mode	Temp (ºC)	Time (M:S)	Acquisition Mode	Cycles
Reverse Transcription					
Step 1.	NA	65	05:00	NA	NA
Step 2.	NA	50	50:00	NA	NA
PCR					
Denaturation	None	95	05:00	None	
Amplification	Quantification	95	00:12	None	45
		60	00:30	Single	

2.3.2 CHIKV and DENV assays

CHIKV and DENV were measured using the clinically validated real-time quantitative RT-PCR (qRT-PCR) protocols routinely used in the diagnostic laboratories (Rare and Important Pathogens Laboratory, PHE). All assays were performed on a LightCycler 480 using the Invitrogen SuperScript™ III One-Step RT-PCR System with Platinum™ Taq DNA Polymerase (Invitrogen). The CHIKV assay targets the E1 gene generating a 127 bp amplicon (*305*) and the two DENV assays target the 3' non-coding region generating a 79 bp amplicon (*306*). The DENV 1-3 assay covers the three equivalent DENV serotypes and is a multiplex with the MS2 PCR, the DENV4 assay covers the remaining DENV serotype. Details on each assay can be found in Table 2.4 and Table 2.5 and the cycling conditions in Table 2.6. The target regions amplified by real-time qRT-PCR were used for the design of custom RNA oligos (Table 2.7) as a standard to estimate copy numbers of detected virus for each sample.

Table 2.4 CHIKV assay reaction mix composition

CHIKV		
Reagent	**Single Reaction**	**Final Concentration**
2x Reaction Mix	10µL	1X
Water	2.215 µL	NA
Forward Primer	0.18 µL	900 nM
Reverse Primer	0.18 µL	900 nM
Probe	0.125 µL	625 nM
MgSO4 (50mM)	1.5 µL	3.75 mM
Platinum Taq	0.8 µL	NA
Template	5 µL	NA
Total	15 µL	NA

Table 2.5 DENV1-3 and DENV4 assay reaction mix composition

Reagent	**Single Reaction**	**Final Concentration**
DENV 1-3		
2x Reaction Mix	10 µL	1x
Water	1.04 µL	NA
Forward Primer	0.06 µL	300 nM
Reverse Primer	0.18 µL	900 nM
Probe	0.1 µL	500 nM
Forward Primer (MS2)	0.08 µL	40 nM
Reverse Primer (MS2)	0.08 µL	40 nM
Probe (MS2)	0.16 µL	80 nM
MgSO4 (50mM)	2.5 µL	6.25 mM
Platinum Taq	0.8 µL	NA
Template	5 µL	NA
Total	15 µL	NA
DENV 4		
2x Reaction Mix	10 µL	1x
Water	1.41 µL	NA
Forward Primer	0.1 µL	50 nM
Reverse Primer	0.09 µL	450 nM
Probe	0.1 µL	500 nM
MgSO4 (50mM)	2.5 µL	6.25 mM
Platinum Taq	0.8 µL	NA
Template	5 µL	NA
Total	15 µL	NA

Table 2.6 CHIKV, DENV1-3 and DENV4 RT-PCR cycling parameters

Step Name	Analysis Mode	Temp (ºC)	Time (M:S)	Acquisition Mode	Cycles	Rate (ºC/sec)
Reverse Transcription	None	45	10:00	None	1	20
Denaturation	None	95	05:00	None	1	20
Amplification	Quantification	95	00:05	None	45	20
		57	00:35	Single		20
Cooling	None	40	00:30	None	1	20

Table 2.7 CHIKV, DENV1-3 and DENV4 custom RNA oligos sequences

Assay	Oligo Sequence	Reference
CHIKV	5'-UCGACGCGCCCUCUUUAACGGACAUGUCGUGCGAGGUACCAGCCUGCACCCAUUCCUCAGACUUUGGGGGCGUCGCCGUUAUUAAAUAUGCAGCCAGUAAGAAAGGCAAGUGUGCGGUGCAUUCGAU-3'	KY751908
DENV 1-3	5'-GGAUAGACCAGAGAUCCUGCUGUCUCCUCAGCAUCAUUCCAGGCACAGAACGCCAGAAAAUGGAAUG-3'	NC_001474
DENV 4	5'-GGAUAGACCAGAGAUCCUGCUGUCUCUGCAGCAUCAUUCCAGGCACAGAGCGGCGCAAGAUGGAUUG-3'	NC_002640

2.3.3 LASV assay

Suspected LASV samples are routinely screened at the Bernhard Nocht Institute for Tropical Medicine and the Institute of Lassa Fever Research and Control (ILFRC) using two real-time qRT-PCR approaches, the commercially available Altona kit (RealStar® Lassa Virus RT-PCR Kit 1.0 CE, Altona Diagnostics, Hamburg, Germany) targeting the S segment along with an in-house version of the previously described Nikisins RT-PCR targeting the L segment (*307*)pileup.

The latter has been optimized by using the SuperScript™ III Platinum™ One-Step qRT-PCR reagents (Invitrogen) according to manufacturer instructions (without magnesium sulfate; reaction volume of 25 µl). The temperature profile was identical to that of the Altona assay, while primer and probe sequences and concentrations were used as described (*307*). Both Altona and Nikisins real-time RT-PCR assays

have been implemented and extensively evaluated in terms of analytical and clinical characteristics by the BNITM, and were found to have good performance when used in combination for detection of the virus in acute Lassa fever cases. Reaction setup can be found in Table 2.8. Both assays are run on the Rotor-Gene Q (Qiagen) and share the same cycling conditions (Table 2.9).

Table 2.8 LASV Altona and Nikisins assay reaction mix composition

Altona		
Reagent	**Single Reaction**	
Master Mix A	5 µL	
Master Mix B	15 µL	
Template	10 µL	
Total	30 µL	
Additional: 1µL Internal Control is added to the positive and negative control		
Nikisins		
Reagent	**Single Reaction**	**Final Concentration**
2x Superscript Mix	12.50 µL	1X
Water	3.95 µL	NA
Forward Primer	1.25 µL	500 nM
Reverse Primer	1.30 µL	520 nM
Probe	0.5 µL	200 nM
Superscript III	0.5 µL	NA
Template	5 µL	NA
Total	25 µL	NA

Table 2.9 LASV Altona and Nikisins RT-PCR cycling parameters

Step Name	Analysis Mode	Temp (ºC)	Time (M:S)	Acquisition Mode	Cycles
Reverse Transcription	None	55	20:00	None	1
Denaturation	None	95	02:00	None	1
Amplification	Quantification	95	00:15	None	45
		55	00:45	Single	
		72	00:15	None	

2.3.5 MS2

MS2 is used as an internal extraction control specifically used in the diagnostic laboratories as an inhibition control for RT-PCR reactions. We utilise MS2 as an internal control for our sequencing runs to validate that the extraction, reverse transcriptase and library preparation has been successful. The MS2 assay targets a 99 bp amplicon using the SuperScript™ III One-Step RT-PCR System with Platinum™ Taq DNA Polymerase (Invitrogen). Reaction setup can be found in Table 2.10 and cycling conditions in Table 2.11.

Table 2.10 MS2 assay reaction mix composition

Reagent	Single Reaction	Final Concentration
2x Reaction Mix	10µL	1X
Water	1.38 µL	NA
Forward Primer	0.08 µL	40 nM
Reverse Primer	0.08 µL	40 nM
Probe	0.16 µL	80 nM
MgSO4 (50mM)	2.5 µL	3.75 mM
Platinum Taq	0.8 µL	NA
Template	5 µL	NA
Total	15 µL	NA

Table 2.11 MS2 RT-PCR cycling parameters

Step Name	Analysis Mode	Temp (ºC)	Time (M:S)	Acquisition Mode	Cycles	Rate (ºC/sec)
Reverse Transcription	None	45	10:00	None	1	20
Denaturation	None	95	05:00	None	1	20
Amplification	Quantification	95	00:05	None	45	20
		57	00:35	Single		20
Cooling	None	40	00:30	None	1	20

2.4 Nucleic acid extraction

Total nucleic acid was extracted from 140 µl of each sample using the QIAamp viral RNA kit (Qiagen) with the addition of linear acrylamide instead of carrier RNA and eluted in 60 µl. All patient samples were inactivated in a biosafety level 3 or 4 laboratory depending on the pathogen classification by addition of AVL buffer (Qiagen) and disinfected with 5,000ppm chlorine solution (HazTabs, Appleton Woods LTD) for 10 min prior to removal and downstream processing in a Biosafety level 2 laboratory.

2.5 DNase treatment and purification

As the extraction kit is not designed to separate viral RNA from host DNA and both nucleic acids will be recovered, thus 30 µl of elute were subjected to DNase treatment with TURBO DNase (Thermo Fisher Scientific) at 37ºC for 30 min to remove any host DNA present. Finally, RNA was purified and concentrated to 8 µl using the RNA Clean & Concentrator™-5 kit (Zymo Research) as per manufacturer's guidelines.

2.6 Agencourt AMPure XP PCR Bead Clean-Up

Agencourt AMPure XP PCR beads (Beckman Coulter Life Sciences) (from herein referred to as AMPure beads) allow for purification and clean-up with no salt carryover. Excess oligos, nucleotides, enzymes and salts are removed resulting in

purified products through selective binding of dsDNA to paramagnetic beads. DNA of 100 bp and larger binds to the beads and is followed by separation of the beads from the solution using a magnetic field. The beads are then washed to remove contaminants and DNA is eluted in the final step from the magnetic beads.

Sample purification is done as per manufacturer's guidelines. Briefly, AMPure beads are stored at 4^{o}C and allowed to come to room temperature prior to use. Beads are mixed well before use until homogenous and 70% Ethanol (EtOH) is freshly prepared. Beads are added to the sample at desired ratio (1:1) and incubated at room temperature for 5 min. The reaction tube is then placed on a DynaMag™-2 Magnet (12321D, Thermo Fisher Scientific) and allowed for beads to separate. Cleared solution is then aspirated and beads are washed 2x with 200 µl of the freshly prepared 70% EtOH. Once the last EtOH wash is removed beads are allowed to air dry before eluting in desired volume.

2.7 Single Primer Isothermal Amplification

The Ovation RNA-Seq System V2 kit (Nugen) - mentioned throughout the thesis as Ribo-SPIA® protocol - was used per manufacturer's instructions for Single Primer Isothermal Amplification (SPIA) of RNA. Briefly, the first step is the generation of the first strand cDNA which is prepared using a first strand DNA/RNA chimeric primer mix and reverse transcriptase. The primers have a DNA portion, comprised of either a random k-mer sequence or polyT, which hybridizer to the 5’ end or randomly across the RNA length to initiate amplification. The reverse transcriptase extends the 3’ DNA end of the primer generating the first strand and the resulting hybrid contains a unique RNA sequence at the 5’ end of the cDNA strand. The final step is the SPIA which uses the DNA/RNA chimeric primer, DNA polymerase and RNase H in a homogeneous isothermal assay which amplifies the cDNA. RNase H allows for the degradation of the RNA in the DNA/RNA duplex which results in exposure of the unique DNA sequence. This provides a binding site for a new DNA/RNA primer and the initiation of replication by DNA polymerase at the 3’ end of the primer displacing the existing forward strand. The RNase H removes the RNA portion at the 5’ end of the newly synthesised cDNA allowing for the next round of cDNA synthesis. The process is repeated resulting in amplification of cDNA.

2.8 Sequence Independent Single Primer Amplification

The SISPA approach was adapted from Greninger *et al (308)*. For the reverse transcription step, 10^4 copies/ml of purified MS2 RNA was included as a control at this step and samples were incubated at 65°C for 5 min with 1 µl (40 pmol/µl) Primer A (5'- GTTTCCCACTGGAGGATA-N9 -3'), followed by the addition of 5 µl SuperScript™ III Reverse Transcriptase (Thermo Fisher) reaction mix (2 µl 5x First-Strand Buffer, 1 µl water, 1 µl 10 mM dNTP mix, 0.5 µl 0.1M DTT, 0.5 µL SSIII RT) and incubation at 42°C for 60 min. Second strand DNA synthesis followed using Sequenase Version 2.0 DNA Polymerase (Thermo Fisher), 5 µl of Sequenase mix #1 (1 µl 5x Sequenase Buffer, 3.85 µl H2O, 0.15 µl Sequenase enzyme) is added to the reaction and incubated at 37°C for 8 min followed by the addition of Sequenase mix #2 (0.45 µl Sequenase Dilution Buffer, 0.15 µl Sequenase Enzyme) and a second incubation at 37°C for 8 min. This concludes the reverse transcription and second strand cDNA synthesis. The final step in the SISPA protocol is the DNA amplification which was performed with AccuTaq™ LA DNA Polymerase (Sigma), in which 5 µl of cDNA synthesised material is added to 45 µl of AccuTaq LA reaction mix (35 µl H2O, 5 µl 10x AccuTaq Buffer, 2.5 µl 12.5 mM dNTP, 1 µl DMSO, 1 µl (100 pmol/µl) Primer B (5′-GTTTCCCACTGGAGGATA-3′), 0.5 µl AccuTaq enzyme) per sample with the following PCR conditions: 98°C for 30s; 30 cycles of 94°C for 15s, 50°C for 20s, and 68°C for 5 min, followed by 68°C for 10 min. Amplified cDNA was purified using a 1:1 ratio of AMPure XP beads (Beckman Coulter, Brea, CA) and stored at -20°C. All samples were quantified when defrosted for the preparation of a sequencing library.

2.9 MiSeq library preparation

MiSeq sequencing libraries were prepared using 1.5 ng of amplified cDNA as input template into the Nextera XT V2 kit (MiSeq, Illumina) and library preparation was conducted as per manufacturer's instructions. Briefly input DNA is tagmented using the Nextera transposome, a process that allows from the simultaneous fragmentation of the DNA and addition of adapter sequences to tag the DNA. The tagmented DNA is then amplified using a limited-cycle PCR which also allows for the addition of index sequences and full adapter sequences to the tagmented DNA. Indexes were selected using the MiSeq experimental manager software and samples were multiplexed in batches (maximum 16 samples per run). Once the amplification step is finished the libraries are cleaned-up using AMPure XP beads (Beckman

Coulter) to purify the DNA and remove short fragments. Quantity of each library is normalised to ensure equal library representation in the pooled library which combines equal volumes of the normalised libraries in a single tube. The pooled library is finally diluted and heat-denatured prior to loading for the sequencing run. Pooled libraries were sequenced on a 2x150 bp paired end Illumina MiSeq run, by the Genomics Services Development Unit, Public Health England, Colindale, London, UK.

2.10 MinION sequencing and library preparation

MinION sequencing libraries were prepared using the latest available consumables and protocols from Oxford Nanopore Technologies at the time of the experiments. Protocols evolved along the three year period of this project. Details on which protocols were used are included in the results chapters but information on all protocols used and a description of the protocol chemistry is included in this section. Whilst all protocols include an optional DNA fragmentation and DNA repair step prior to the main protocol, neither of these steps were included in any of the library preparations used. The MinION device is connected to a computer through a USB port and configured through the MinKNOW™ Software Agent, which was updated prior to each experiment to ensure the most recent version of the software was used.

MinION sequencing libraries were prepared using the total available amplified cDNA of each sample with a maximum input of 1000 ng, in the majority of samples processed this level was not reached, thus input was less than recommended, details are specified in each results chapter. All steps of the MinION protocol were performed in DNA LoBind tube (Eppendorf). Prior to any sequencing run, the flow cell used was quality controlled by running the Platform QC script on MinKNOW, assessing the number of active pores in the flow cell, generally a total of 450 or more active pores in Group 1 was considered acceptable to continue with the experiment. Once quality checked the flow cell was primed using the Flow Cell priming mix prior to loading the prepared library, which was sequenced using the 48hr script for all samples. Copies of the MinION library preparation protocols, which are briefly described below, can be found in the Supplementary material section (Protocols S1-S4).

2.10.1 2D DNA by ligation (SQK-NSK007 and SQK-LSK208)

The 2D DNA by ligation protocol (SQK-NSK007 and SQK-LSK208, Oxford Nanopore Technologies) has now been discontinued but was used for experiments in Chapter 4. The protocol allowed sequencing of both strands within a dsDNA molecule using a leader adaptor and a hairpin adaptor, which allowed for both the template and complement strands of a dsDNA to pass through a pore leading to a 2D read. An overview of the protocols is shown in Figure 2.1.

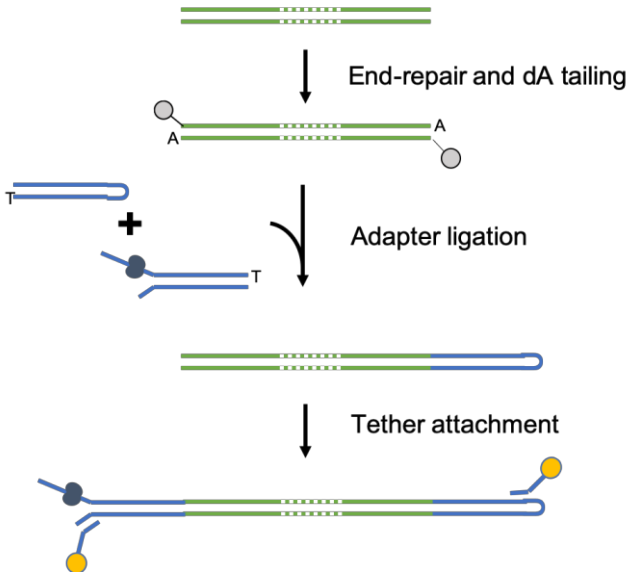


Figure 2.1 2D DNA by ligation protocol overview

(Figure adapted from https://community.nanoporetech.com/)

Briefly, input DNA was end-repaired and dA-tailed using the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs) and incubated for 5 min at 20°C and 5 min at 65°C using a heat block. DNA was purified using 60 µl AMPure beads and eluted in 31 µl nuclease free water. End-prepped/dA-tailed DNA was quantified with an expected recovery of ~700ng and 30 µl was taken forward for adapter ligation. The adaptor and the hair pin adaptor were added using Blunt/TA Ligase Master Mix (New England Biolabs) followed by the addition of the tether. Dynabeads™ MyOne™ Streptavidin C1 (Thermo fisher Scientific) were used to purify the adapted and tethered DNA library, which was then eluted in 15 µl of elution buffer and quantified. The final library was prepared, loaded and the sequencing run was initiated. At the time of these experiments, the Metrichor™ Desktop Agent managed the connection to the base-calling service in the cloud hosted by Metrichor, which was run in parallel to the sequencing run generating basecalled fast5 files.

2.10.2 1D^2 kit (SQK-LSK308)

The 1D^2 kit (SQK-LSK308, Oxford Nanopore Technologies) chemistry utilises the same principles as the 1D kit with the main difference being the sequencing adaptors, which in this case are designed to encourage the complementary strand to immediately be sequenced after the template. This protocol in combination with appropriate 1D^2 data analysis produces reads with higher accuracy than the 1D protocol. The protocol incorporates ligation of the 1D^2 adapters onto the end-repaired/dA-tailed DNA fragments prior to the ligation of the sequencing adaptors and the tether. An overview of the protocol is shown in Figure 2.2.

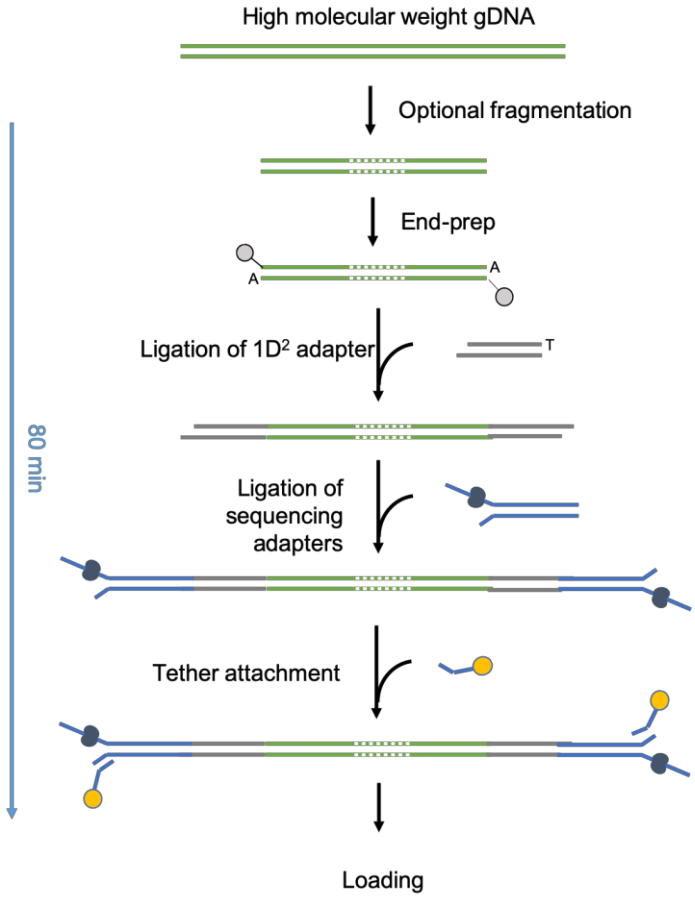


Figure 2.2 1D^2 DNA by ligation protocol overview

(Figure adapted from https://community.nanoporetech.com/)

Briefly, input DNA was end-repaired and dA-tailed using the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs) and incubated for 5 min at 20°C and 5 min at 65°C using a heat block. The reaction was then cleaned-up using 60 µl AMPure beads, eluted in 23 µl nuclease free water and quantified.

Sequencing 1D^2 adaptors were added using the Blunt/TA Ligase Master Mix (New England Biolabs) and 20 µl AMPure beads were added to purify the adapter ligated DNA, which was eluted in 46 µl of elution buffer and quantified. Sequencing adaptors were then added using the Blunt/TA Ligase Master Mix (New England Biolabs) and 40 µl AMPure beads were added to purify the adapter ligated DNA, which was washed using the Adapter Bead Binding Buffer, eluted in 15 µl of elution buffer and quantified. Library was prepared, loaded using Library Loading Beads (EXP-LLB001) and the sequencing run was initiated.

2.10.3 Rapid kit (SQK-RAD003)

The Rapid kit (SQK-RAD003, Oxford Nanopore Technologies) utilises a transposase-method for the sequencing library preparation during which a transposome complexes (transposase in combination with two adapters) are used to fragment DNA molecules and simultaneously add adaptors to the fragmented DNA (used in Chapter 4). The approach is characterised by its 10min library preparation. An overview of the protocols is shown in Figure 2.3.

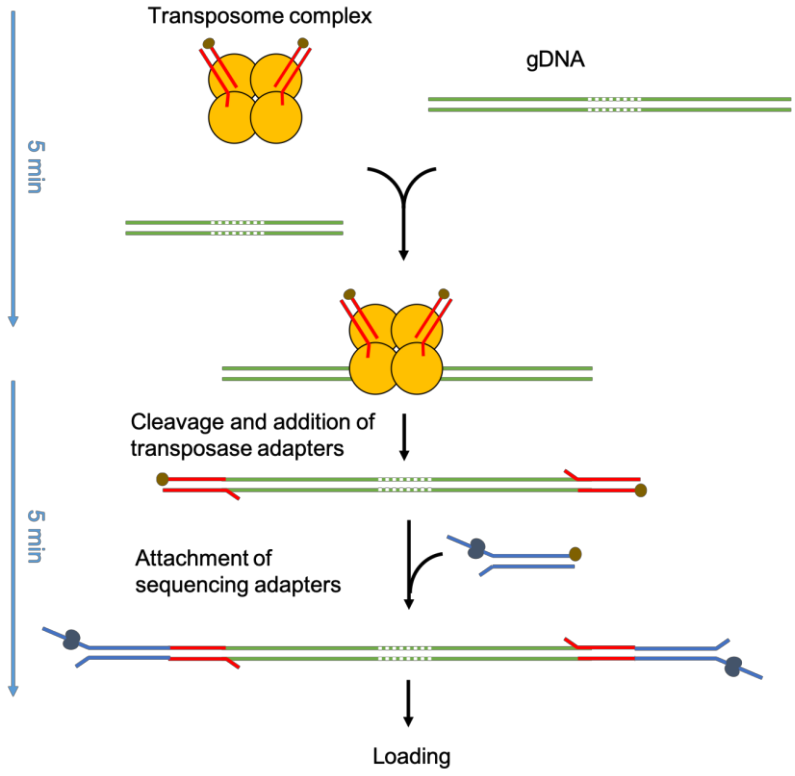


Figure 2.3 Rapid Sequencing protocol overview

(Figure adapted from https://community.nanoporetech.com/)

Briefly, input DNA of approx. 400 ng, in a volume of 7.5 µl, was mixed with 2.5 µl Fragmentation Mix (FRA) and incubated for 1 min at 30°C and 1 min at 80°C. Following the fragmentation step, 1 µl of Rapid adaptor (RPD) was added and incubated for 5 min at room temperature. Library was prepared, loaded using Library Loading Beads (EXP-LLB001) and the sequencing run was initiated.

2.10.4 1D DNA by ligation (SQK-LSK108)

The 1D protocol (Oxford Nanopore Technologies) allows for sequencing of DNA by the addition of a leader adaptor ligated onto each end of a dsDNA and is currently one of the most widely used kits (used in Chapter 5). The leader adaptor consists of two partially complementary oligos that form a Y-shaped structure when annealed. A tether is added to the adaptors during library preparation, which anchors the DNA fragments to the membrane, in close proximity to the nanopore facilitating binding and sequencing initiation. The leader adaptor with the motor protein binds to the nanopore and the single strand of that leader adapter is fed through the nanopore. Once the strand, referred to as the template, is sequenced the complement strand dissociates and possibly gets sequenced as a separate event.

An overview of the protocol is shown in Figure 2.4. Input DNA was end-repaired and dA-tailed using the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs) and incubated for 5 min at 20°C and 5 min at 65°C using a heat block. The reaction was then cleaned-up using 60 µl AMPure beads and eluted in 31 µl nuclease free water. End-prepped/dA-tailed DNA was quantified and 30 µl was taken forward for adapter ligation. Adaptor ligation used Blunt/TA Ligase Master Mix (New England Biolabs) with a 10min incubation and 40 µl AMPure beads were added to purify the adapter ligated DNA, which was washed using the Adapter Bead Binding Buffer, eluted in 15 µl of elution buffer and quantified. Library was prepared, loaded using Library Loading Beads (EXP-LLB001) and the sequencing run was initiated.

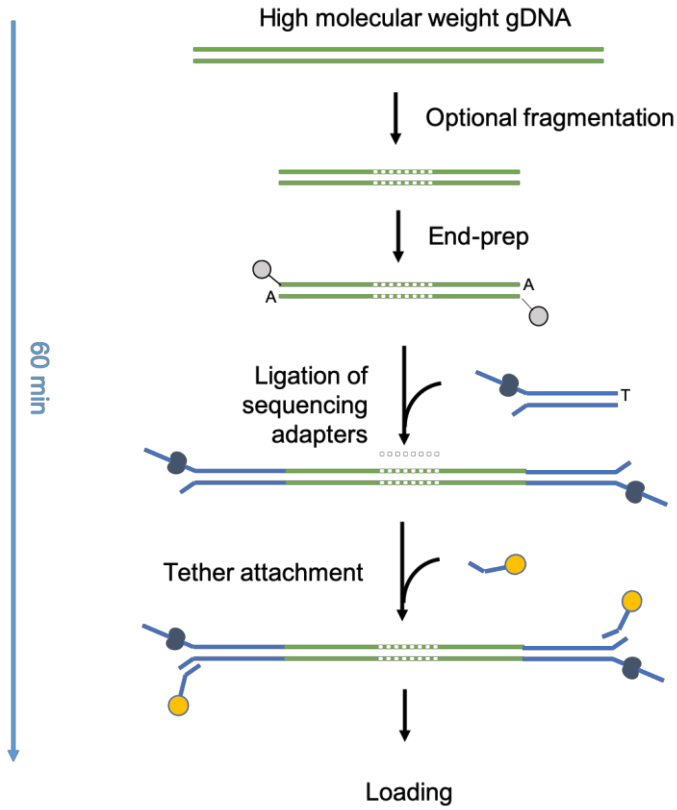


Figure 2.4 1D DNA by ligation protocol overview

(Figure adapted from https://community.nanoporetech.com/)

2.10.5 Barcoding Kit (EXP-NBD103)

Barcoded MinION sequencing libraries were prepared using the Ligation sequencing kit 1D (SQK-LSK108, Oxford Nanopore Technologies) and the Native Barcoding Kit (EXP-NBD103, Oxford Nanopore Technologies) (Used in Chapter 5). PCR-free multiplexing of samples is achieved using 12 unique barcodes. The barcodes are added at each end of the dsDNA by ligation prior to the addition of the leader adaptors and the tether.

An overview of the protocol is shown in Figure 2.5. Briefly, input DNA was end-repaired and dA-tailed using the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs) and incubated for 5 min at 20°C and 5 min at 65°C using a heat block. The reaction was then cleaned-up using 60 µl AMPure beads and eluted in 23 µl nuclease free water. A unique barcode was selected per sample, end-prepped/dA-tailed DNA was quantified and 500 ng of each sample were taken forward for barcode ligation. Each reaction was then cleaned-up using 50 µl AMPure beads and eluted in 26 µl nuclease free water. Each barcoded DNA sample was quantified and equimolar amounts of each were pooled together producing a pooled

sample of 700 ng in total in 50 µl. Sequencing adaptors were added to the pooled barcoded sample using the Quick Ligation™ Kit (New England Biolabs) and 40 µl AMPure beads were added to purify the adapter ligated DNA, which was washed using the Adapter Bead Binding Buffer, eluted in 15 µl of elution buffer and quantified. Library was prepared, loaded using Library Loading Beads (EXP-LLB001) and the sequencing run was initiated.

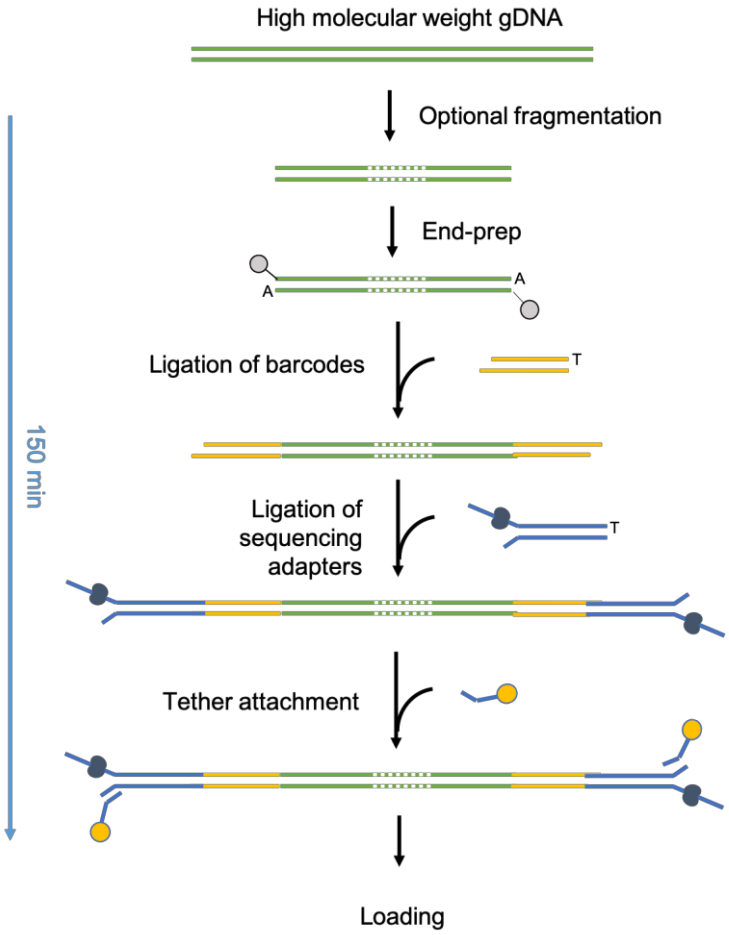


Figure 2.5 Barcoding Kit coupled with 1D DNA by ligation protocol overview

(Figure adapted from https://community.nanoporetech.com/)

2.11 Data analysis

Two main routes were used to analyse sequencing data. The first one is reference mapping using an existing sequence, to which the reads are aligned. The second is *de novo* assembly, which is the reference free assembly of contiguous sequences from the reads.

2.12 Reference assisted alignment consensus

2.12.1 Mapping and alignment statistics

The Burrows-Wheeler Alignment tool (BWA) (*309*) was used to align reads to a reference sequence, details on specific version and reference sequences used are included in the results chapters. BWA allows for mapping of low-divergence sequences against a reference genome and consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. We utilise BWA-MEM, which is designed to run with longer sequences (70 bp to 1 Mbp), is faster, more accurate and recommended for high-quality queries (*310*). BWA-MEM is based on an algorithm that identifies super-maximal exact matches originally published in 2012 (*311*) and later extended into its current form in 2013 to allow for BWA-MEM to be a fully featured mapper (*309*, *312*). Samtools, is a suite of utilities for manipulating sequencing alignments. Bedtools, is a suite of programs for working with SAM/BAM, BED and VCF files. Available online resources were utilised for all tools (*313*–*317*).

The first step is for BWA to construct the FM-index for the reference genome used. Sub-command to index the reference is:

```
bwa index ref.fasta
```

Once the reference is indexed BWA-MEM is used to align the query sequences to the reference. BWA-MEM is also able to process MinION long reads and account for their accuracy using the -x option and specifying ont2d. Command for alignment of MiSeq paired-end reads to a reference and for alignment of MinION reads to a reference:

MiSeq:

```
bwa mem -t 8 ref.fasta sample.R1.fastq sample.R2.fastq > sample.sam
```

MinION:

```
bwa mem -x ont2d -t 8 ref.fasta sample.fastq > sample.sam
```

BWA output file format is a Sequence Alignment/Map format (SAM). SAM is a TAB-delimited format consisting of an optional header and an alignment section. A SAM file needs to be transformed to its binary counterpart the BAM format, for further analysis. This is done using samtools with the options -S (specifies that the input file is a SAM file) and -b (indicates that the output file will be in BAM format), the two requirements can be merged and represented as -Sb in the command. The alignment produced is in a random order with respect to their position in the reference genome, leading to a BAM file, which includes the information in the order the sequences occur in the input fastq files.

```
samtools view -Sb sample.sam > sample.bam
```

Certain analysis tools require that the BAM file is sorted so that the alignment information is present in the reference genome order, i.e. the information needs to be ordered based on their alignment coordinates to the reference genome.

```
samtools sort sample.bam -o sample.sort.bam
```

Samtools is designed to run as part of a stream thus takes into account the input file as stdin and the output file as stdout allowing for several commands to be combined using UNIX pipes. An example of this functionality is included below for a piped analysis of MiSeq data.

```
bwa mem -t 8 ref.fasta sample.R1.fastq sample.R2.fastq | samtools view -Sb - | samtools sort -o
sample.ref.illumina.bam
```

Next step is to index the sorted BAM file which allows for quick extraction of alignment information and is required by genome viewers (e.g. Tablet (*318*), The James Hutton University) to quickly display alignment when navigating to different regions of the genome. The Samtools command to index the sorted bam file will create an additional index file in the current directory where the command is run.

```
samtools index sample.sort.bam
```

Samtools and bedtools were used to generate additional information. Samtools was used to generate Samtools flagstat that goes through the input file to calculate and print statistics to stdout and this is then passed and written in a text file to store the information.

```
samtools flagstat sample.sort.bam > sample.flagstat.txt
```

Samtools depth computes the depth at each location. To output all positions including positions with depth zero the -a option needs to be included in the command.

```
Samtools depth -a  –d 100000000 sample.bam > sample.depth.txt
```

Bedtools genomecov computes histograms and can take a BAM file as input (-ibam). The histogram's maximum can be controlled using the -max option.

```
bedtools genomecov -ibam sample.sort.bam -max 20 -g > sample.cov20.txt
```

Commands were put together in a shell script. All required files (sample.fastq and reference.fasta) were always located in the parent directory and all generated files were saved in the current directory where the script was run from. Commands were included in a text file and saved as a .sh file. An example of the MiSeq script and how to run using the command line:

```
#command to run script: source scriptname.sh sample_file_name_prefix reference_name

#sample_file_name_prefix($prefix)=$1
#reference_name($ref)=$2

bwa index ../$ref.fasta

bwa mem -t 8 ../$ref.fasta ../$prefix.R1.fastq.gz ../$prefix.R2.fastq.gz | samtools view -Sb - |
samtools sort -o $prefix.to.$ref.illumina.bam

samtools index $prefix.to.$ref.illumina.bam

samtools flagstat $prefix.to.$ref.illumina.bam > $prefix.to.$ref.illumina.flagstat.txt

Samtools depth -a  –d 100000000 $prefix.to.$ref.illumina.bam > $prefix.to.$ref.illumina.depth.txt

bedtools genomecov -ibam $prefix.to.$ref.illumina.bam -max 20 -g >
$prefix.to.$ref.illumina.cov20.txt
```

2.12.2 Variant calling and consensus generation

For MiSeq sequencing data, quasibam (*319*), an in house developed C++ program, was used to process the BAM file and generate the consensus sequence. Quasibam uses a pileup based method and generates a majority consensus along with a table of nucleotide frequency, depth and quality metrics for each nucleotide position in the mapping process. Minimum variation for inclusion in consensus was set at 20, inferring any mixture greater than 20% to be coded as IUPAC ambiguities. Minimum number of reads required to report consensus position was set at 20. All other parameters were run as default, setting options listed in Table 2.12.

Table 2.12 Quasibam parameters and options used

Parameter	Option
mapq minimum	30
Minimum percentage of variants reported	1
Minimum variation for inclusion in consensus sequence	20
ASCI string cutoff for fastq	33
Minimum number of reads required to report a consensus position	20
Insert optimisation?	Yes please
Remove PCR duplicates?	Yes please
Perform error checking based on position of minority variants in a read?	Yes please
Fraction at each end of a read where errors are expected.	0.1
Fraction of reads containing a variant that are within the read region above.	0.9
Report fasta header as key value pairs?	No thanks

For MinION data, Nanopolish and margin_cons script were used in combination. Nanopolish (*320*) allows for the calculation of an improved consensus sequence for a draft genome assembly, detects base modifications, calls SNPs and indels with respect to a reference genome. Nanopolish was utilised as previously described by Quick et. al. (*147*). Briefly, this reference mapping approach utilises nanopolish variants to align signal-level events and detect variants in regards to the most closely related reference genome; insertions and deletions were not called due to the nature of the sequencing data. Nanopolish required access to the signal-level data (fast5) output from the nanopore sequencer. Nanopolish options used can be found in Table 2.13. This leads to the first step of the nanopolish workflow, which is the indexing of the fast5 files by telling nanopolish where to locate the files. Providing the sequencing_summary.txt file speeds up the indexing process.

```
nanopolish index -d /fast5_files -s sequencing_summary.txt albacore_output.fastq
```

Once the fast5 files are indexed using nanopolish, bwa is used to index the reference genome and to align the basecalled reads to the reference sequence. Samtools is used to sort the resulting bam file and the consensus algorithm is run using nanopolish variants, which computes the vcf file with the variants information.

Finally margin_cons.py script (*218*), which is integrated in the zika pipeline of the ZIBRA project (*321*), was used to generate the consensus from the vcf file generated from nanopolish and the reference used as input in nanopolish. The commands were put together in a script:

```
#Run command: source scriptname.sh refname readsname sampleid segment runnumber
#$1=reference name
#$2=reads name
#$3=isth_id
#$4=segment
#$5=RUNNUMBER xxx

bwa index ../References/$1.fasta

bwa mem -x ont2d -t 10 ../References/$1.fasta ../seqtk/$2.fastq | samtools sort -o
Sorted.$3.$1.$4.bam -T readsmap.$1.$4.tmp -

samtools index Sorted.$3.$1.$4.bam

nanopolish variants -t 10 --ploidy 1 --snps -r ../seqtk/$2.fastq -b Sorted.$3.$1.$4.bam -g
../References/$1.fasta -o $3.$4.$1.NPSNPS.vcf --min-candidate-frequency 0.1 2>&1 | tee -a
logfile.$2.$1.$4.log

python /zika-seq-master/pipeline/scripts/margin_cons.py ../References/$1.fasta
$3.$4.$1.NPSNPS.vcf Sorted.$3.$1.$4.bam > $3.$4.NpSnpsMrgCons.Consensus.fasta 2>
logfileMarginCons.$2.$1.$4.log
```

Table 2.13 Nanopolish parameters and options used

Parameter	**Option**
variants	Find SNPs using a signal-level HMM
-t	Number of threads used
--ploidy	Ploidy level of the sequenced genome
--snps	Only call SNPs
-r	Reads in fasta format
-g	Reference genome (fasta)
-o	Output file
--min-candidate-frequency	Extract candidate variants from the aligned reads when the variant frequency is at least F (default 0.2)

Additionally, for MinION data, a simple pileup script (Figure S1) with bases called at a minimum depth of 20x and 70% support fraction was used, either independently or in succession to Nanopolish and margin_cons script. The pileup

script generates a majority consensus along with a table of nucleotide frequency and depth metrics for each nucleotide position in the mapping process. Any base location that does not fulfil the depth and support fraction is assigned an N IUPAC ambiguity code.

2.13 *De novo* assembly

MiSeq generated sequences were assembled *de novo* using Spades 3.8.2 (*322*) in combination with SSPACE Standard v3.0 (*323*). Spades is a short-read genome assembly module and SSPACE Standard is a script based on SSAKE that allows for scaffolding of pre-assembled contigs. Both have been designed to run with paired-end libraries. The Spades command and the parameters (Table 2.14) used are as follows:

```
python ../../spades.py --pe1-1 sample.R1.fastq.gz --pe1-2 sample.R2.fastq.gz --cov-cutoff 5 --careful -t 6 -k 21,33,55,77,99,127 -o Sample_Spades_OutputFolder
```

Table 2.14 Spades parameters and options used

Parameter	Option
--pe1-1	File with left reads for paired-end library
--pe1-2	Files with right reads for paired-end library
--cov-cutoff	Read coverage cut-off value.
--careful	Minimises number of mismatches in the final contigs
--only-assembler	Option not included, this is for reads that have been corrected prior to running assembler
-t	Number of threads
-k	K-mer lengths, increased in increments of 22 until k-mer length reaches 127
-o	Output directory

The contig.fasta output from SPades is then used as input for SSPACE, which is able to order, distance and orientate the contigs and combine them into scaffolds. SSPACE requires a library file, which contains the paired read files and information on the insert size, error and orientation. The library file provides information on the

appropriate format for the reads to be stored (e.g. paired reads are stored in a new file with similar read names for easy identification of the paired reads) and once the reads are formatted provides information on mapping the filtered paired reads. The format of the information included in the library.txt file is as follows and the information included in each column of the file is in Table 2.15.

```
Lib1 bwa ./Sample.R1.fastq.gz ./Sample.R2.fastq.gz 250 0.25 FR
```

Table 2.15 SSPACE column information

Column	Option
1	Name of library
2	Name of aligner to be used for library
3 and 4	Fasta or fastq files - paired end reads
5	Expected insert size
6	Maximum allowed error
7	Orientation of the paired- reads

Once the library file is ready and is located in the directory where SSPACE will be run, SSPACE is run with the following command:

```
perl ./SSPACE_Standard_v3.0.pl -l libraries.txt -s ./Sample_Spades_OutputFolder/scaffolds.fasta
```

MinION sequencing data was assembled *de novo* using CANU (*324*), a fork of the Celera Assembler designed to work with high-noise long-read sequences such as the Oxford Nanopore MinION output. Canu has three functions which can run independently or combined: correction, trimming and assembly. Canu has novel features allowing for improved performance with high noise long read data such as auto-detection of computational resources to scale itself to the available resources, adaptive k-mer weighting and automated error rate estimation. Canu was run with the following command, all functions combined and parameters as mentioned in Table 2.16.

```
canu -d Sample_Parameters -p Sample_Parameters -nanopore-raw sample.fastq
genomeSize=10000 minReadLength=400 minOverlapLength=200 corOutCoverage=10000
gnuplotTested=true
```

Table 2.16 CANU parameter information

Parameter	Option
-d	Allows canu to create an assembly-directory and run in that directory
-p	Set the file name of intermediate and output files
-nanopore-raw	Reads option to describe how reads were generated
genomeSize	Estimate of the size of the target genome. Is used to decide how many reads will be corrected and how sensitive the mhap overlapper should be. Also impacts some logging, in particular reports the NG50 sizes.
minReadLength	Reads shorter that the input value will not be loaded in the assembler. Reads output by correction and trimming that are shorter than this will also be discarded.
minOverlapLength	Overlaps shorter than this will not be taken into account. Smaller values are used to overcome lack of read coverage, but also lead to false overlaps and potential assemblies. The higher the value the more accurate assemblies but more fragmented too.
corOutCoverage	Only correct the longest reads up to this coverage, defaults to 40

2.14 Metagenomic data analysis

Metagenomic laboratory preparation of samples is important to allow for identification of pathogens that are not previously known to be present. To allow for this it is important to utilise the correct tools to analyse this data thus to allow for non-targeted screening of the data. Samples often contain multiple pathogens, pathogens that cross react in the diagnostic assays used, or new pathogens that were not considered previously to have caused human disease leading to difficulty in identifying all these scenarios unless the right tools are used.

Bioinformatics tools for metagenomic analysis are designed to methodically screen sequencing data and return classified reads within a list of species. Metagenomic classification tools have been used to characterise the composition of a variety of samples including clinical patient samples.

Metagenomic analysis was conducted using the taxonomic classifier Kraken v0.10.4-beta (*325*) and Centrifuge v1.0.4-beta (*326*). For Kraken Taxonomic labels

were assigned against a locally built database populated with bacterial, viral, archaeal genomes and a representative yeast genome. The majority of sequences included in the database are complete genomes from the RefSeq database (v.66) with an addition of 141 viral GenBank sequences including several Hepatitis E viruses, Parechovirus and HAZV. For Centrifuge taxonomic classification was performed against the provided Bacterial, Archaea, Viruses, Human (compressed) indexes (last updated: 12/06/2016).

2.15 Basecalling

Basecalling was performed in the first sequencing runs using Metrichor (Oxford Nanopore Technologies) and Albacore v1.2 (Oxford Nanopore Technologies) and later only using Albacore v2.2.7 (Oxford Nanopore Technologies). Metrichor was a cloud-based basecalling approach that is no longer available, as it was replaced with Albacore which performed local basecalling. Albacore is a data processing pipeline that provides the basecalling algorithms and post-processing options for the fast5 files generated by the MinION. As of Albacore v1.1 and later the algorithm could basecall directly to fastq files using the option -o fastq. The command used to basecall a run using the latest Albacore version used was:

```
read_fast5_basecaller.py -f Flowcell -k Kit -o OutputFileType -i Input_Data -s Output_Directory -t #_Threads -r
```

2.16 Porechop

Porechop (*327*) is a tool for identifying and removing barcodes and sequencing adaptors from Oxford Nanopore reads. Sequencing barcodes and adapters are identified at the ends of reads and trimmed, additionally reads that were barcoded are demultiplexed. Porechop is able to identify reads with adapters in the middle, which are considered chimeric and separated to two reads. Porechop v0.2.3 was used to trimm and demultiplex reads directly from the Albacore output directory using the command:

```
porechop -i input_reads.fastq -b output_dir
```

2.17 SeqTK

SeqTK (*328*) is a fast tool for processing FASTA and FASTQ files. It allows for fast manipulation of both files. Command used to convert fastq files to fasta:

```
seqtk seq -a in.fq.gz > out.fa
```

Command used to trim 30bp from the left and 30bp from the right end of each read:

```
seqtk trimfq -b 30 -e 30 input.fastq > output.fastq
```

2.18 Samtools fastq

Samtools fastq allows for the conversion of a BAM file into a fastq. The -f option is used to only output alignments with all bits set in INT present in the FLAG field. In this case -f 4 is used which outputs all unmapped reads. This command was used to generate fastq files that excluded any reads mapping to human. An other example of using samtools fastq is to extract all reads that mapped, in which case we would use -F 4. The -F option does not output alignments with any bits set in INT present in the FLAG field, so it would not include any unmapped reads.

```
samtools fastq -f 4 sample.bam > sample_unmapped.fastq
```

2.19 Awk and Bioawk

The awk (*329*) language allows users to select records within a file and perform operations on them. Bioawk (*330*) is an extension to awk adding specific support for multiple sequencing and biological data formats including SAM, FASTA and FASTQ. Commands sourced online have been used to manipulate fastq files. Read length and frequency of identified read length was extrapolated using:

```
awk 'NR%4 ==  2 {lengths[length($0)]++} END {for (l in lengths) {print l, lengths[l]}}'  file.fastq
```

Commands using bioawk to count the number of total sequences in a fastq file and creation of a tab delimited table with names and sequence length:

```
bioawk -cfastx 'END{print NR}' test.fastq

bioawk -c fastx '{ print $name, length($seq) }' < Sample.fasta > Sample_ReadNames&Length.txt
```

2.20 Chapter 3 specific material and methods

2.20.1 Sample selection

The HAZV mock-sample was prepared at a concentration of 10^6 viral copies/ml using a stock of culture grown HAZV and sterile-filtered human serum (S7023, Sigma-Aldrich). Subsequently four positive routine diagnostic samples, two plasma and two serum, were obtained from RIPL, Public Health England (PHE), Porton Down. All had previously tested positive by real-time qRT-PCR for DENV with different cycle threshold (Ct) values, one low (Ct: 17.67), two medium (Ct: 24. and 25.28) and one high (Ct: 28.69). RT-PCR assays were used for confirmation and quantitation of DENV (Chapter 2, section 2.3.2).

2.20.2 Sample preparation and sequencing

Samples were extracted (Section 2.4), DNase treated and purified (Section 2.5) and AMPure XP beads were used for clean-up (Section 2.6). The Ovation RNA-Seq System V2 kit (Nugen) - referred to throughout the thesis as Ribo-SPIA® protocol - was used as per manufacturer's instructions for SPIA of RNA (Section 2.7). SISPA protocol was carried out as described in Section 2.8. Illumina libraries of all samples were prepared as per Section 2.9.

2.20.3 Data handling

References used for alignments in this chapter were (Genbank ID): HAZV L segment (NC_038709.1), HAZV M segment (NC_038710.1), HAZV S segment (NC_038711.1), DENV Serotype 1 (NC_001477.1), DENV Serotype 2 (NC_001474.2), DENV Serotype 3 (NC_001475.2), DENV Serotype 4 (NC_002640.1). Reads were randomly normalised using fastq_pair (*331*) and the -n option (number of read subselection) was set based on the total number of Ribo-

SPIA® reads generated per sample. Further details on the data analysis and the tools used can be found in Chapter 2.

2.21 Chapter 4 specific material and methods

2.21.1 Sample selection

Twenty-six positive routine diagnostic samples, nine plasma and 17 serum, were obtained from the RIPL, PHE, Porton Down. All had previously tested positive by real-time qRT-PCR for CHIKV or DENV, with a maximum cut-off value of Ct 35. Drosten et al. (*306*) and Edwards et al (*305*). RT-PCR assays were used for confirmation and quantitation of DENV and CHIKV respectively, as described in Section 2.3. Samples were selected based on their Ct values, among a larger set of 441 samples, so as to represent a Ct clinical range.

2.21.2 MinION library preparation and sequencing

MinION sequencing libraries were prepared using total amplified cDNA of each sample to a maximum of 1 µg. Oxford Nanopore kits SQK-NSK007 or SQK-LSK208 (2D), SQK-LSK308 ($1D^2$) and SQK-RBK001 (Rapid) were used and each sample was run individually on the appropriate flow cell (FLO-MIN105, FLO-MIN106 or FLO-MIN107) using the 48hr run script. Base calling was performed using Metrichor (Oxford Nanopore Technologies) for SQK-NSK007 and SQK-LSK208 or Albacore v1.2 for SQK-LSK308 and SQK-RBK001. Poretools (*332*) was used to extract FASTQ files from Metrichor FAST5 files.

2.21.3 Data handling

BWA-MEM v0.7.15 (*309*) was used to align reads to the following references (Genbank ID): DENV Serotype 1 (NC_001477.1), DENV Serotype 2 (NC_001474.2), DENV Serotype 3 (NC_001475.2), DENV Serotype 4 (NC_002640.1) and CHIKV (NC_004162.2). Mapping consensuses for MiSeq were generated using in-house software QuasiBam (*319*) and for MinION using a simple pileup. Nanopolish variants (*147*, *320*) was used in consensus mode to compute error-corrected consensus sequence for the Rapid kit. De novo assemblies for MinION sequences were run using CANU v1.6 (*333*, *334*) with the following settings: corOutCoverage=1000, genomeSize=12000, minReadLength=300, minOverlapLength=50.

2.22 Chapter 5 specific material and methods

2.22.1 Sample Collection

Samples from suspected Lassa fever patients were routinely tested for presence of LASV RNA at the ILFRC, Irrua Specialist Teaching Hospital (ISTH), Irrua, Edo State, Nigeria, using two real-time reverse transcription PCR assays (Section 2.2.3); the commercially available Altona kit (RealStar® Lassa Virus RT-PCR Kit 1.0 CE, Altona Diagnostics, Hamburg, Germany) targeting the S segment and the Nikisins RT-PCR targeting the L segment (*307*). A total of 120 plasma, breast milk, or cerebrospinal fluid samples identified as LASV-positive by one or both real-time PCRs were selected based on temporal and geographical spread across the outbreak.

2.22.2 MinION Library Preparation and Sequencing

Barcoded MinION sequencing libraries were prepared using the Ligation sequencing kit 1D (SQK-LSK108) and Native Barcoding Kit (EXP-NBD103) (Section 2.10.4 and 2.10.5). Up to 6 samples plus one negative control (consisting of a water blank sample included in each batch of extractions) were included per multiplex library. Reads generated from the negative controls were taken through the analysis pipeline with no significant viral assemblies generated and no viral sequence produced. If the negative controls were identified positive for LASV with a viable viral sequence of more than 50% consensus sequence of a segment, the complete run would be dismissed and repeated. Libraries were sequenced for 48hr on FLO-MIN106 flow cells using a Mark 1B MinION device.

2.22.3 Data Handling

Reads generated were basecalled using Oxford Nanopore Technologies Albacore software v2.2.7 (Section 2.15) and basecalled fastq files were concatenated and demultiplexed using Porechop (Section 2.16). SeqTK was used to trim from both ends (Section 2.17) to eliminate primer sequences and resulting fastq files were mapped to the human genome, human_g1k_v37 (1000 Genomes) and mapped reads were excluded from the subsequent *de novo* assembly. Owing to the diversity of LASV, selection of an individual reference genome for read alignment was required for each sample. CANU v1.6 (Section 2.13) was used for *de novo* assembly, which

allowed for LASV reference identification. CANU genomeSize and minReadLength parameters were lowered for samples that did not assemble any LASV contigs with the specified values. Assemblies were used in a blastn search against the NCBI database to identify the closest LASV reference genome available and BWA was used for read alignment (Section 2.12) to the reference genome identified. Nanopolish variants and margin_cons.py script (*321*) were used to filter out low-quality or low-coverage candidate SNPS and compute the final consensus respectively (Section 2.12.2). Samtools was used to compute percentage reads mapped along with coverage depth and bedtools was used to calculate genome coverage at 20x (Section 2.12). Taxonomic classification of the data was performed using Centrifuge (Section 2.14).

2.22.4 Hardware equipment

Four nanopore MinION devices, and the computer hardware necessary were put together. The hardware used (Figure 2.6) included 3x MacBook Pro laptop, along with a custom-build computer (Table 2.17) built in a light and compact (260 x 208 x 280 mm) Cool Master case, running Ubuntu 14.4. The devices were set-up with all expected requirements for software. Additionally four Samsung T5 2TB USB 3.0 & USB-C (Samsung) external solid state drives and two WD Elements USB 3.0 4TB (Western Digital) portable external drives were used to allow for fast data manipulation and back-ups, subset of each were formatted for use with the MacBook (Mac OS Extended) and the remaining formatted for use with Linux computer (EXT4).

Figure 2.6 Hardware equipment

Table 2.17 Computer Specifications of custom build desktop computer.

iItem	**Catalog Number**	**Manufacturer**
Cooler Master Elite 110A	RC-110A-KKN1	Cool Master
Intel - Core i9 - 7980XE 2.6GHz 18-Core Processor	i9-7980XE	Intel
Cooler Master - MasterLiquid Lite 120 66.7 CFM Liquid CPU Cooler	MLW-D12M-A20PW-R1	Cooler Master
ASRock - X299E-ITX/ac Mini ITX LGA2066 Motherboard	X299E-ITX/ac	ASRock
SanDisk - X400 1TB M.2-2280 Solid State Drive	SD8SN8U-1T00-1122	SanDisk
Corsair - HX Platinum 750W 80+ Platinum Certified Fully-Modular ATX Power Supply	HX750 Platinum	Corsair
Corsair - Vengeance 32GB (2 x 16GB) DDR4-3000 Memory	CMSX32GX4M2A3000C16	Corsair
Western Digital - Black 1TB 2.5" 7200RPM Internal Hard Drive	WD10JPLX	Western Digital

Chapter 3

Method evaluation for metagenomic sequencing of RNA viruses

3. Chapter 3. Method evaluation for metagenomic sequencing of RNA viruses

3.1 Overview

Two methods of unbiased cDNA amplification for viral metagenomic sequencing were compared: The Ribo-SPIA® approach, used with Illumina sequencing for viral metagenomic sequencing in the 2014 EBOV outbreak (*144*) and a SISPA protocol, which had been previously successfully coupled with the MinION for detection of viral pathogens in clinical samples (*308*). Samples, including a mock sample spiked with HAZV and four DENV clinical samples, were prepared using both protocols, sequenced on an Illumina MiSeq and assessed based on percentage of reads mapping to the viral genome and consensus sequence recovery at 10x and 20x depth. The SISPA protocol outperformed the Ribo-SPIA® particularly for the DENV clinical samples, leading to higher percentages of total reads mapping to the virus and near-complete genomes for all four DENV samples.

3.2 Introduction

Metagenomic protocols require no *a priori* knowledge of the pathogen of interest for their design, offering a hypothesis-free approach to target identification. The aim of this chapter was to identify possible methods for metagenomic sequencing of RNA viruses on the MinION platform. From the pre-existing literature, two leading methods were identified for further evaluation, Ribo-SPIA® and SISPA.

Ribo-SPIA® had been extensively used on clinical samples during the 2014 EBOV outbreak (144, 336) and was a well established method available, but had not been tested coupled with nanopore sequencing. Ribo-SPIA® uses a single-primer isothermal linear amplification and has been shown to generate full genomes of EBOV, human immunodeficiency virus, respiratory syncytial virus and WNV from ultra-low copy viral samples, as low as 100 copies of viral RNA (335). SISPA had been demonstrated to work with nanopore sequencing, however the number of clinical samples tested was low. SISPA uses tagged random primers for the amplification of cDNA and has been extensively used for virus detection and full genome sequencing of viruses (see Section 1.6 Sequence-independent single primer amplification).

Sequence-independent sequencing using RNA single primer isothermal amplification (Ribo-SPIA®, Ovation® RNA-Seq System V2, NuGEN), previously mainly used for transcriptome analysis, was successfully applied by Gire et al. (*144*), Malboeuf *et al.* (*335*) and Lewandowski *et al.* (*336*) for the identification of RNA viruses in clinical samples. Ribo-SPIA® utilises a mixture of random and oligo (dT) DNA/RNA chimeric primers for the reverse transcription step, generating the first strand cDNA which serves as a template for the second strand synthesis. For the second strand synthesis, the RNA strand within the DNA/RNA duplex is fragmented creating a priming site for the DNA polymerase to synthesize the second strand resulting in a double stranded cDNA with a unique DNA/RNA heteroduplex at one end. The final step is the single-primer isothermal amplification which uses the DNA/RNA chimeric primer, DNA polymerase and RNase H in a homogeneous isothermal assay which amplifies the cDNA (Figure 3.1B).

Metagenomic sequencing of CHIKV, EBOV and Hepatitis C was demonstrated in principle using a SISPA protocol coupled with the MinION by Greninger *et al.* in 2015 reporting their detection from a small number of human blood samples (*308*). The SISPA approach utilizes a primer consisting of a random 3' end sequence, more specifically nine random N (each N position can be an A, C, T or G nucleotide) in the protocol used by Greninger *et al. (308)*, and a defined sequence tag at the 5' end. The tagged nonamer is used for the reverse transcription of the extracted RNA, followed by the second strand synthesis to generate dsDNA which is subsequently amplified using the defined sequence as a primer-binding extension sequence (Figure 3.1A).

To compare the two protocols and assess their feasibility and sensitivity, both methods were used for the non-specific RNA amplification of a mock sample spiked with HAZV and of four DENV positive clinical samples with a variety of Ct values. Both protocols were evaluated based on the proportion of viral nucleic acid recovered from each sample and on the percentage recovery of complete genome sequences for each virus.

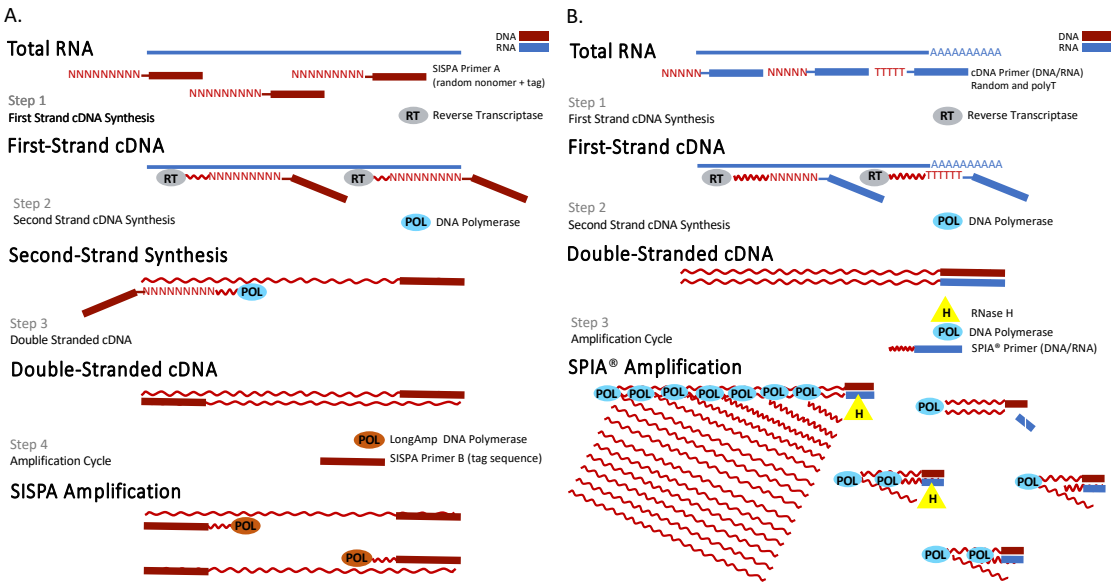


Figure 3.1 Schematic diagram of SISPA and Ribo-SPIA® amplification processes.
(A) SISPA protocol outline. SISPA primer A consisting of a random 3' end sequence (nine random N and a defined sequence tag at the 5' end) is used for the reverse transcription of the extracted RNA, followed by the second strand synthesis to generate dsDNA which is subsequently amplified using a long amplification DNA polymerase and SISPA primer B (the tag sequence of SISPA primer A) as a primer-binding extension sequence. (B) Ribo-SPIA® protocol outline. A mixture of random and oligo(dT) DNA/RNA chimeric primers are used for the reverse transcription step. For the second strand synthesis, the RNA strand within the DNA/RNA duplex is fragmented creating a priming site for the DNA polymerase. A single-primer isothermal amplification using the DNA/RNA chimeric primer, DNA polymerase and RNase H in a homogeneous isothermal assay allows for cDNA amplification. (Section B of the figure was adapted from: Ovation RNA Amplification System V2 User guide PART NO. 3100-12, 3100-60, 3100-A01)

3.3 Results

3.3.1 HAZV spiked sample assessment

A serum sample spiked with HAZV at a titre of 10^6 genome copies per ml was sequenced using both SISPA and Ribo-SPIA® approaches. The proportion of viral nucleic acid present relative to host/background and the genome coverage achieved can be seen in Table 3.1. The proportion of total reads mapping to each segment (L: large segment, M: medium segment, S: small segment) of the virus was substantial using either metagenomic protocol. The percentage of total reads mapping to the complete HAZV genome was 59% for the Ribo-SPIA® protocol and 78% for the SISPA protocol. Using the Ribo-SPIA® protocol 35%, 11% and 3.7% mapped to HAZV segments L, M and S respectively and for SISPA protocol the equivalent percentages were 56%, 14% and 9% (Figure 3.2). Irrespective of lower mapping percentages observed with the Ribo-SPIA® protocol both approaches produced sufficient viral reads to achieve genome coverage of >98% for each of the three segments (Figure 3.2). The SISPA approach produced a slightly lower percentage of genome coverage both at 10x and 20x, with 70 bp less coverage at 10x (SISPA: 99.16%, Ribo-SPIA®: 99.96%) and 85 bp less at 20x, a negligible difference in coverage of the complete genome (18232 bp) of HAZV (Table 3.1).

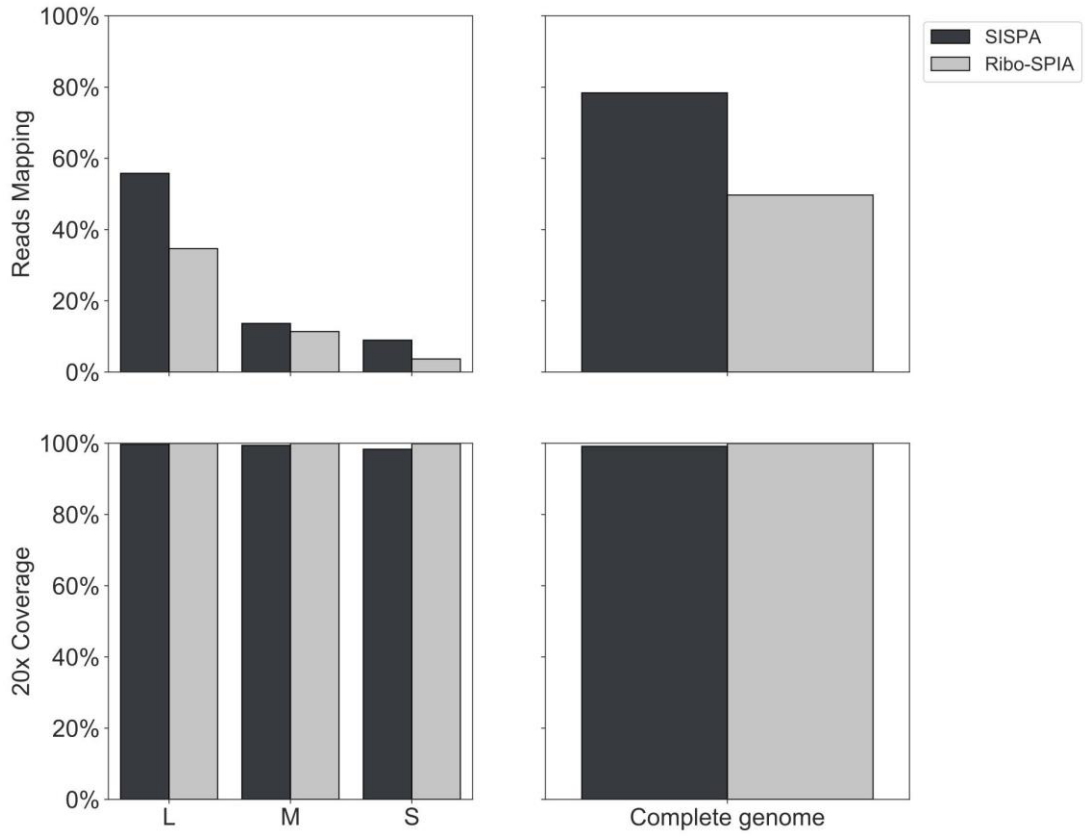


Figure 3.2 Proportion of reads mapping to the appropriate viral reference sequence and proportion of reference genome sequenced at minimum 20-fold coverage in the HAZV mock sample.

The percentage of total reads mapping to each segment reference sequence and the complete genome is plotted in the upper panel. Lower panels display the percentage of the reference genome sequenced to a minimum depth of 20-fold in the Illumina data.

Table 3.1 Description of sequencing mapping data to HAZV for the SISPA and Ribo-SPIA® prepared samples,

Sample (method)	Total Reads (R1+R2)[a]	Reads mapping to HAZV	% Reads mapping to HAZV	10x depth	20x depth	Reference accession	Ref size (nts)
Complete Genome							
SISPA	2288576	1794072	78,39 %	99.40 %	99.16 %	NC_038709.1-NC_038711.1	18232
Ribo-SPIA	1260234	626222	49,69 %	100%	99.96 %	NC_038709.1-NC_038711.1	18232
L Segment							
SISPA	2288576	1276899	55,79 %	99.74 %	99.73 %	NC_038711.1	11980
Ribo-SPIA	1260234	436671	34,64 %	100%	100%	NC_038711.1	11980
M Segment							
SISPA	2288576	312268	13,64 %	99.54 %	99.41 %	NC_038710.1	4575
Ribo-SPIA	1260234	143252	11,36 %	100%	100%	NC_038710.1	4575
S Segment							
SISPA	2288576	204905	8,95 %	98.92 %	98.33 %	NC_038709.1	1677
Ribo-SPIA	1260234	46299	3,67 %	100%	99.88 %	NC_038709.1	1677

[a] 'R1+R2' indicates paired end sequencing

Figure 3.3 shows the percentage of reads mapping to the human genome reference and specifically to human ribosomal sequences. The Ribo-SPIA® protocol yielded a higher percentage of reads mapping to both with 33.70% and 6.25% total reads mapping to the complete human genome and ribosomal sequences respectively, compared to the SISPA protocol which are lower; 17.22% to human and 0.15% to ribosomal references (Table 3.2).

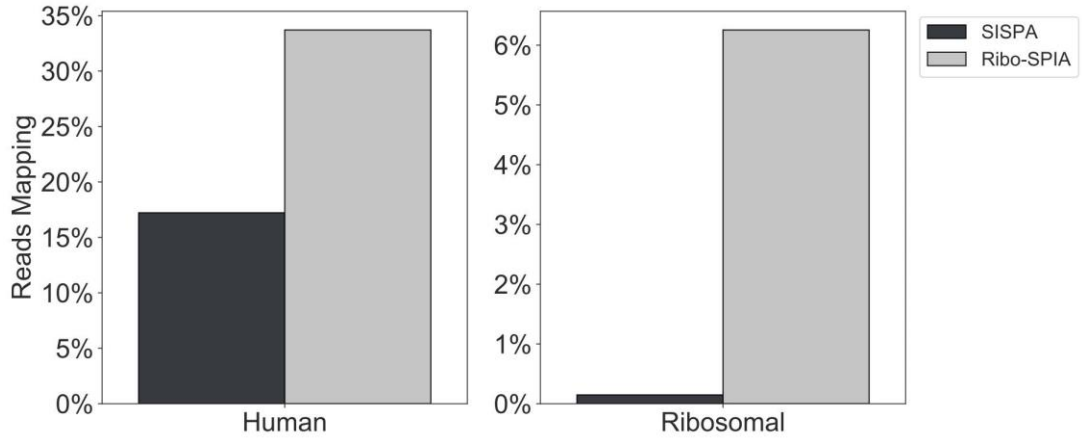


Figure 3.3 Proportion of reads mapping to the human genome and human ribosomal sequences.

The percentage of total reads mapping to each appropriate reference sequence with read percentage mapping to human shown on the left and to ribosomal on the right.

Table 3.2 Description of sequencing mapping data to human and ribosomal sequences for the SISPA and Ribo-SPIA® prepared samples.

Sample (method)	Total reads (R1+R2) [a]	Reads mapping to human	% Reads mapping to human	Reads mapping to ribosomal	% Reads mapping to ribosomal
SISPA	2288576	394105	17.22 %	3369	0.15 %
Ribo-SPIA	1260234	424817	33.70 %	78817	6.25 %

[a] 'R1+R2' indicates paired end sequencing

The coverage depth of reads mapped across the segments for each approach is shown in Figure 3.4, the actual depth is a function of the total reads sequenced as the read levels are not normalised. A 20x depth per location is achieved in both cases for all three segments.

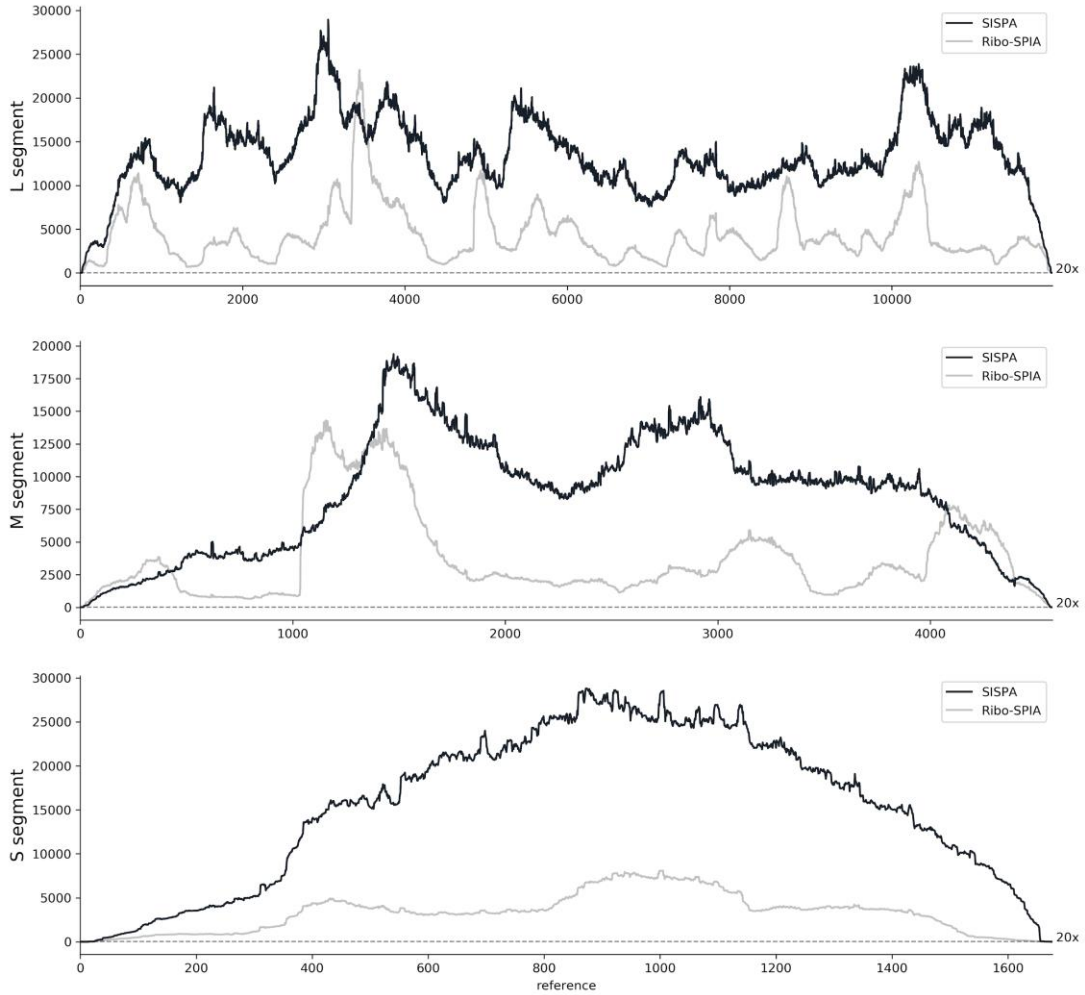


Figure 3.4 Coverage depth across the HAZV viral genome, (n = 2 samples).
Each graph corresponds to a HAZV segment. Read depth (y-axis) across the genome (x-axis) following reference alignment is shown. SISPA coverage is shown in black and Ribo-SPIA® coverage is indicated in grey. Total depth has not been normalised; comparison is to show overall pattern of coverage. Dotted horizontal line indicates depth of 20x coverage, used for consensus calling.

3.3.2 Clinical sample assessment

To further evaluate and compare the two protocols using genuine clinical samples, four DENV positive clinical samples (two serum and two plasma) were selected based on real-time qRT-PCR Ct value from samples identified positive in RIPL diagnostic laboratories, PHE, Porton Down. Samples were selected to include one low (DENV 3, Ct: 17.76), two medium (DENV 8, Ct: 24.8 and DENV 9, Ct: 25.28) and one high (DENV11, Ct: 28.69) Ct value samples. DENV samples selected were prepared using both the SISPA and Ribo-SPIA® protocols and sequenced. The proportion of reads mapping to the respective DENV viral reference was significant

for all four samples (Figure 3.5, Table 3.3, Table 3.4), with the lowest proportions of reads mapping observed in the samples prepared with the Ribo-SPIA® approach.

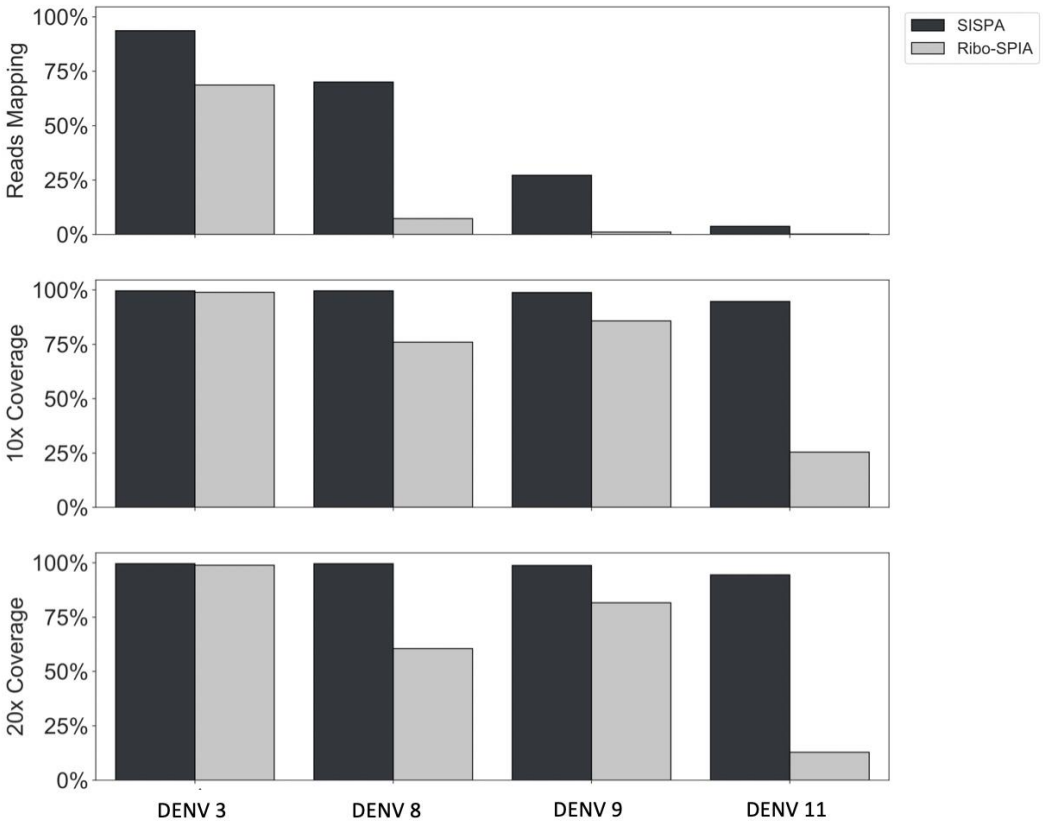


Figure 3.5 Comparison of SISPA and Ribo-SPIA® results, as to proportions of reads mapping to the appropriate reference viral sequence, and proportion of reference genome sequenced at minimum 20-fold and 20-fold coverage (n = 4 samples).

The percentage of total reads mapping to the appropriate reference sequence is plotted in the upper panel. The middle panel displays the percentage of reference genome sequenced to a minimum depth of 10-fold in the data generated and the lower panel displays the percentage of the reference genome sequenced to a minimum depth of 20-fold in the data generated.

Table 3.3 Description of samples positive for DENV with corresponding reference sequences used for mapping.

Sample	Ct value	Sample type	Reference virus[a]	Reference accession	Reference size (nts)
DENV 3	17.67	Plasma	DENV 2	NC_001474.2	10723
DENV 8	24.8	Serum	DENV 3	NC_001475.2	10707
DENV 9	25.28	Plasma	DENV 2	NC_001474.2	10723
DENV11	28.69	Serum	DENV 1	NC_001477.1	10735

[a]DENV serotype is also indicated

Table 3.4 Description of samples positive for DENV with corresponding Illumina mapping data for SISPA and Ribo-SPIA .

Sample	Ct value	Method sequenced	Total reads (R1+R2)[a]	Total reads mapping	%Reads mapping	% 10x coverage	% 20x coverage
DENV 3	17.67	SISPA	738814	691805	93.64 %	99.59 %	99.59 %
		Ribo-SPIA	397712	273045	68.65 %	98.92 %	98.89 %
DENV 8	24.8	SISPA	777264	544315	70.03 %	99.59 %	99.56 %
		Ribo-SPIA	90116	6565	7.29 %	76.02 %	60.49 %
DENV 9	25.28	SISPA	787728	214347	27.21 %	98.82 %	98.78 %
		Ribo-SPIA	670376	7435	1.11 %	85.78 %	81.65 %
DENV 11	28.69	SISPA	1034698	38641	3.73 %	94.71 %	94.45 %
		Ribo-SPIA	459622	943	0.21 %	25.47 %	12.87 %

[a]'R1+R2' indicates paired-end sequencing

The read proportion per sample showed good concordance with clinical Ct values. Sample DENV3 with the lowest Ct value (Ct: 17.67) had the highest mapped read percentages observed for both protocols used, 93.63% with SISPA and 68.65% with Ribo-SPIA®. In sample DENV8 and DENV9 with mid-Ct values (Ct: 24.8 and Ct: 25.28 respectively), the percentage of reads mapping was lower, 70% and 27.21% with the SISPA protocol and 7.29% and 1.11% with the Ribo-SPIA® protocol respectively. Whilst in the high Ct sample (DENV11, Ct: 28.69) the viral proportion drop to 3.73% with SISPA and 0.21% with Ribo-SPIA® protocol.

Figure 3.5 and Table 3.3 show the percentages of reads mapping to the appropriate viral reference along with the 10-fold and 20-fold genome coverage percentage. Using the Ribo-SPIA® approach only DENV3 (Ct: 17.67) had a genome coverage over 90% for both 10x and 20x genome coverage. Using the Ribo-SPIA®

protocol, DENV9 (Ct: 25.28) had a genome coverage of 85.78% at 10x and 81.65% at 20x, followed by DENV8 (Ct: 24.8) which had 76.92% and 60.49% respectively; the lowest genome coverage observed was for DENV11 (Ct: 28.69) with 25.46% at 10x and 12.87% at 20x. On the contrary, all of the SISPA sequenced DENVsamples returned over 90% genome coverage at 10x and 20x, irrespective of lower mapping percentages at the mid and high Ct value samples (DENV9 and DENV11).

Figure 3.6 and Table 3.5 shows the percentages of reads mapping to the human genome reference and to human ribosomal sequences for each DENV sample and for each protocol. The Ribo-SPIA® protocol yielded a higher percentage of reads mapping to both whole human genome and ribosomal sequence across all four samples with the higher Ct samples presenting higher percentages mapped for both. When mapping to the human genome the lowest percentage observed was for sample DENV3 (Ct:17.67), 20,18% with the Ribo-SPIA® approach 1.65% with the SISPA approach. The highest percentage for each approach was 80.32% for Ribo-SPIA® and 56.39% for SISPA, for samples DENV9 (Ct: 25.28) and DENV11 (Ct: 28.69) respectively. The lowest percentage of total reads mapping to ribosomal sequences was 14.8% with the Ribo-SPIA® protocol and 0.75% with the SISPA, both resulting from sample DENV3 (Ct: 17.67). The highest percentage observed for the Ribo-SPIA® protocol was 92% and for the SISPA protocol 49%, both for sample DENV9 (Ct: 25.28).

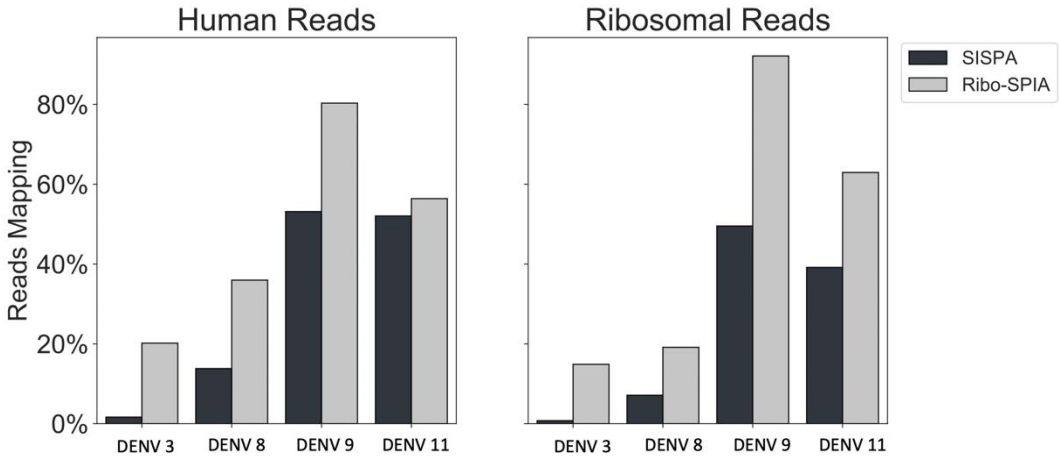


Figure 3.6 Proportion of reads mapping to the human genome and human ribosomal sequences in each DENV positive sample (n = 4 samples).

The percentage of total reads mapping to each appropriate reference sequence with read percentage mapping to human shown on the left and to ribosomal on the right.

Table 3.5 Description of sequencing mapping data to human and ribosomal sequences for the SISPA and Ribo-SPIA® prepared samples of each DENV positive sample sequenced, (n= 4 samples)

Sample	Ct value	Method sequenced	Total reads (R1+R2)[a]	Total reads mapping to human	% Reads mapping to human	Total reads mapping to ribosomal	% Reads mapping to ribosomal
DENV 3	17.67	SISPA	738814	12238	1.66 %	5554	0.75 %
		Ribo-SPIA	397712	80269	20.18 %	59145	14.87 %
DENV 8	24.8	SISPA	777264	107400	13.82 %	55422	7.13 %
		Ribo-SPIA	90116	32417	35.97 %	17249	19.14 %
DENV 9	25.28	SISPA	787728	418665	53.15 %	390173	49.53 %
		Ribo-SPIA	670376	538501	80.33 %	617661	92.14 %
DENV 11	28.69	SISPA	1034698	538529	52.05 %	405188	39.16 %
		Ribo-SPIA	459622	259164	56.39 %	289307	62.94 %

[a]'R1+R2' indicates paired-end sequencing

Figure 3.7 shows coverage depth of reads mapped across the relevant genome for each sample sequence by both the Ribo-SPIA® and SISPA protocol. Read levels are normalised thus depth is a function of the same number of total reads sequenced.

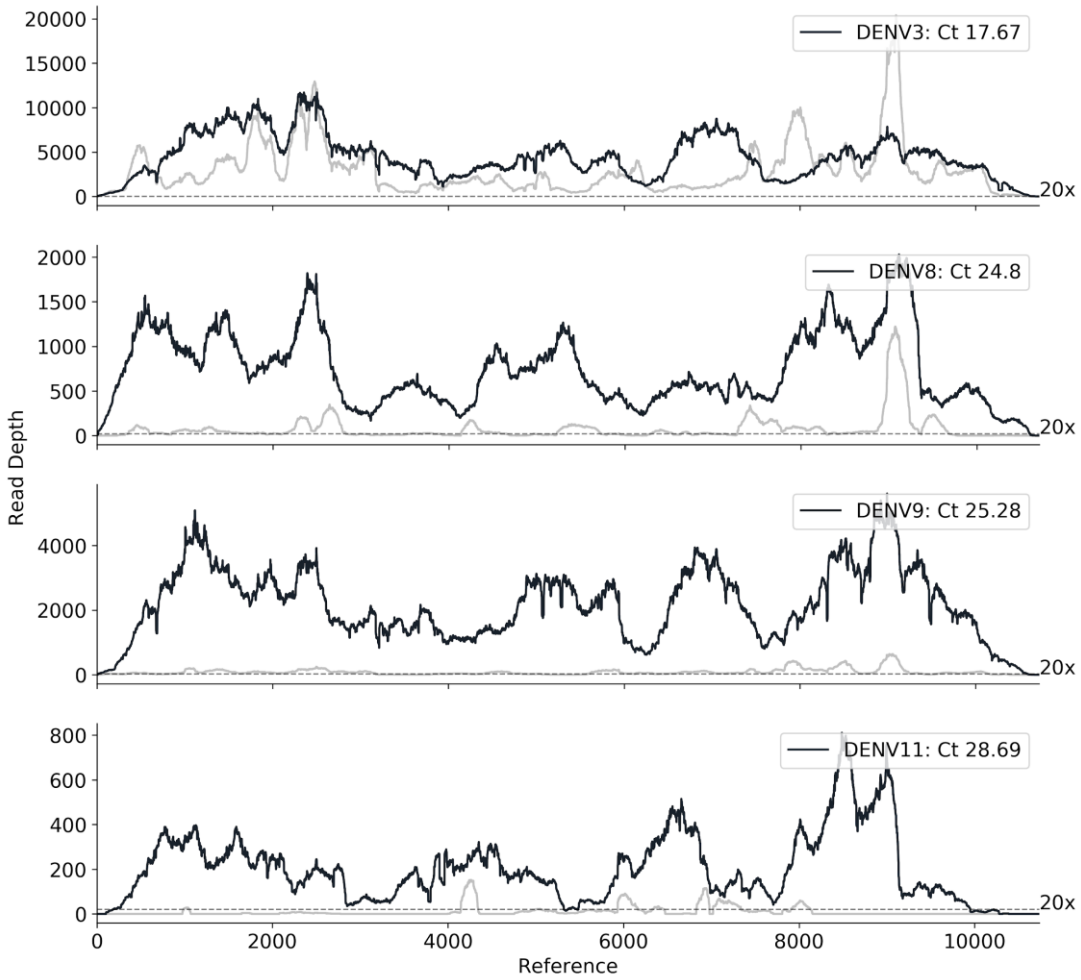


Figure 3.7 Coverage depth across the DENV genome (n = 4 samples).
Each graph corresponds to a given sample defined by its Ct value. Read depth (y-axis) across the genome (x-axis) following reference alignment is shown. SISPA coverage is shown in black and Ribo-SPIA® coverage is indicated in grey. Total depth has been normalised; comparison is to show the comparison in depth of coverage between the two methods. Dotted horizontal line indicates depth of 20x coverage, used for consensus calling.

As seen in Table 3.5 all Ribo-SPIA® prepared DENV samples generated a lower total number of reads compared to the SISPA prepared DENV samples, thus SISPA generated reads were down-sampled randomly using fastq-sample (*337*) to create a corresponding subset normalised to match the number of total reads generated by the equivalent Ribo-SPIA® prepared sample (Table 3.6). The SISPA protocol percentage of reads mapping to the virus is significantly higher than the Ribo-

SPIA® protocol and despite the down-size of the SISPA total read number the 20-fold depth of the genome is >90% for all four DENV samples (Figure 3.7, Table 3.6). For DENV3 (Ct:17.67) both protocols generated comparable 20-fold genome coverage, 98.89% with Ribo-SPIA® and 99.51% with SISPA. For the remaining three DENV samples the Ribo-SPIA® 20-fold genome coverage drops to 60.49% (DENV8, Ct: 24.8), 81.65% (DENV9, Ct: 25.28) and 12.87% (DENV11, Ct: 28.69). Contrary to the Ribo-SPIA® drop in genome coverage at 20-fold the percentage of genome coverage for the SISPA remained above 90% for all DENV samples: 98.91% (DENV8, Ct: 24.8), 98.71% (DENV9, Ct: 25.28) and 90.56% (DENV11, Ct: 28.69).

Table 3.6 Description of samples positive for DENV with corresponding Illumina mapping data post read normalisation

Sample	Ct value	Method sequenced	Total reads (R1+R2)[a]	Total reads mapping	%Reads mapping	% 20x coverage	Reference virus[b]	Reference accession	Reference size (nts)
DENV 3	17.67	SISPA	397712	372473	93.65 %	99.51 %	DENV 2	NC_001474.2	10723
		Ribo-SPIA	397712	273045	68.65 %	98.89 %	DENV 2	NC_001474.2	10723
DENV 8	24.8	SISPA	90116	62914	60.81 %	98.91 %	DENV 3	NC_001475.2	10707
		Ribo-SPIA	90116	6565	7.29 %	60.49 %	DENV 3	NC_001475.2	10707
DENV 9	25.28	SISPA	670376	182399	27.21 %	98.71 %	DENV 2	NC_001474.2	10723
		Ribo-SPIA	670376	7435	1.11 %	81.65 %	DENV 2	NC_001474.2	10723
DENV 11	28.69	SISPA	459622	17011	3.70 %	90.56 %	DENV 1	NC_001477.1	10735
		Ribo-SPIA	459622	943	0.21 %	12.87 %	DENV 1	NC_001477.1	10735

[a]'R1+R2' indicates paired-end sequencing

3.4 Discussion

Sequencing of the HAZV mock sample using the SISPA and Ribo-SPIA® protocol yielded comparable results and despite their difference in total reads mapping to the virus reference sequences both approaches results in >98% genome coverage at 20-fold depth for all three segments demonstrating the promising potential of the SISPA protocol for metagenomic sequencing of RNA viruses. A higher percentage of total reads mapping to the viral reference (Figure 3.1, Table 3.1) and a lower percentage of reads mapping to human and ribosomal sequences (Figure 3.2, Table 3.2) was observed for the SISPA protocol, explained by the abundance of oligo(dT) primers in the first strand cDNA synthesis from template RNA of the Ribo-SPIA® protocol. Nevertheless, both protocols were successful in generating sufficient viral reads for the recovery of complete/near-complete viral genome sequences (Figure 3.1, Table 3.1) and good coverage across the viral segments (Figure 3.3).

The additional evaluation of the two protocols using four DENV positive clinical samples demonstrated that the SISPA protocol outperformed the Ribo-SPIA® method, both in percentage of total reads mapping and genome consensus coverage of the virus of interest. An increase in Ct value coincided with a decreased proportion of viral reads in both protocols, however the percentage of SISPA-generated viral reads were substantially higher compared to the Ribo-SPIA® ones. The percentage of viral reads obtained by the Ribo-SPIA® protocol were 68.65% (DENV3, Ct:17.67), 7.29% (DENV8, Ct:24.8), 1.11% (DENV9, Ct:25.28) and 0.21% (DENV11, Ct:28.69), compared to 93.64%, 70.03%, 27.21% and 3.73% respectively with the SISPA approach. A difference in total percentage of reads mapping to the virus was expected as it was also observed with the HAZV mock sample, although it was not anticipated to be as significant as was observed. The difference between the two methods is further highlighted by the depth and coverage of genomic information recovered for each sample with each approach. The Ribo-SPIA® protocol achieved 20-fold genome coverage of 98.89% (DENV3, Ct:17.67), 60.49% (DENV8, Ct:24.8), 81.65% (DENV9, Ct:25.28) and 12.87% (DENV11, Ct:28.69), which is notably less compared to the SISPA genome coverage (DENV3: 99.59%, DENV8: 99.56%, DENV9: 98.78%, DENV11: 94.45%) particularly for sample DENV8 and DENV11. As expected DENV3 with the lowest Ct value (Ct: 17.67) had the highest percentage of reads mapping to DENV for both protocols and DENV11 with the highest Ct value (Ct: 28.69) had the lowest percentage of reads mapping to DENV. Despite the lower percentage of reads

mapping to the virus, the number of reads was sufficient to succeed a 20-fold genome coverage of >90%.

To determine the background nucleic acid levels present in each sample, reads were mapped to the human genome and to ribosomal sequences; percentages of reads mapping to each were higher for samples sequenced with the Ribo-SPIA® approach. The lowest percentages were observed for DENV3 (Ct: 17.67) which had 20.18% reads mapping to human and 14.87% to ribosomal with the Ribo-SPIA® protocol, compared to 1.66% and 0.75% respectively with the SISPA protocol. Much higher percentages were observed for samples DENV9 and DENV11 particularly with the Ribo-SPIA® protocol, for example from the total DENV9 Ribo-SPIA® reads generated 80.33% mapped to the human genome and 92.14% to ribosomal sequences, notably the respective percentages from the total DENV9 SISPA reads were 53.15% and 49.53%.

The Ribo-SPIA® protocol has been shown to generate full-length genomes from clinical samples for HIV, respiratory syncytial virus, West Nile virus (*335*) and for three EBOV samples (*336*). However, despite these previous findings it was not as successful in generating complete viral genomes for all four DENV samples investigated. The high yield of viral sequences from clinical DENV samples using the SISPA protocol along with the near-complete genome sequences generated for all four DENV samples with different Ct values and three different DENV serotypes (DENV 3: DENV 2, DENV 8: DENV 3, DENV 9: DENV 2 and DENV 11: DENV 1) highlights the benefits and potential of a metagenomic approach to elucidate genomic sequences of RNA viruses directly from clinical samples. Evaluation of the two protocols on DENV clinical samples demonstrates that the total percentage of viral reads and the percentage of viral genome recovery is higher with the SISPA protocol making it a promising approach for metagenomic MinION viral whole-genome-sequencing.

3.5 Conclusions

The initial comparison of the two methods using the HAZV mock sample yielded comparable results and despite a difference in total reads mapping to the virus, sufficient viral reads were generated to obtain near-complete viral genomes for both methods. The subsequent sequencing of the four DENV diagnostic samples allowed for the comparison of the two protocols using genuine clinically relevant samples. The SISPA protocol outperformed the Ribo-SPIA®, leading to higher percentages of total reads mapping to the virus and near-complete genomes (>95%

coverage at 20-fold) for all four DENV samples. The SISPA protocol results demonstrate its unbiased sequencing potential for pathogen identification and complete viral genome recovery and its previous successful coupling with the MinION for viral detection (*308*) makes it particularly promising for rapid portable metagenomic sequencing. To investigate the SISPA protocol capabilities further and determine its sensitivity it was taken forward for further testing using a selection of CHIKV and DENV positive clinical samples positive with a representative range of viral titres.

Chapter 4

Metagenomic Sequencing for Clinical Sample Investigation: Assessing the range of sequencing feasibility for Chikungunya and Dengue Virus

4. Chapter 4. Metagenomic Sequencing for Clinical Sample Investigation: Assessing the range of sequencing feasibility for Chikungunya and Dengue Virus

4.1 Overview

To investigate the feasibility of metagenomic sequencing for recovering whole genome sequences of CHIKV and DENV from clinical samples, a total of 26 samples with a representative range of clinical Ct values were tested. Direct metagenomic sequencing of nucleic acid extracts from serum and plasma without viral enrichment was performed using both the Illumina MiSeq and the portable Oxford Nanopore MinION. The approach successfully allowed for virus and coinfection identification, subtype determination and in the majority of cases elucidated complete or near-complete genomes adequate for phylogenetic analysis. It was demonstrated that metagenomic whole genome sequencing is feasible for the majority of CHIKV and DENV PCR-positive patient serum/plasma samples and the use of MinION metagenomic sequencing was successful for both viruses, highlighting the applicability of this approach to front-line public health and potential portable applications utilizing the MinION.

4.2 Introduction

Human disease outbreaks caused by arboviruses have increased in prevalence over recent decades; led by the spread of mosquito-borne arboviruses such as YFV, CHIKV, DENV, WNV, and ZIKV viruses across both hemispheres (2). CHIKV and DENV are of particular global health concern, as they have lost the need for enzootic amplification and consequently have caused extensive epidemics (3). Their recent global emergence causes significant human disease and their common vectors, symptoms and geographical distribution make their diagnosis both important and challenging. Circulation of CHIKV and DENV (and other arboviruses) in the same areas leads to challenges in differential diagnosis, especially in endemic regions in which diagnosis is predominantly symptom-based and in cases of arboviral coinfections, reports of which have increased in recent years (*32*). Genome sequencing using a metagenomic approach could prove useful for their differentiation

as it has the potential to both identify a viral pathogen present in a clinical sample and provide genomic level data for downstream analysis.

Metagenomic sequencing of CHIKV was demonstrated in principle on the MinION by Greninger et al. in 2015 reporting the detection of CHIKV from a human blood sample (28). Additionally, metagenomics identified CHIKV coinfections within a ZIKV sample cohort (29), with a high proportion of CHIKV reads present making it a promising target for the approach. To test the feasibility of direct metagenomic sequencing for both DENV and CHIKV genomes across a representative range of viral loads a cohort of clinical serum and plasma samples across a representative range of viral loads were selected. The objective was to assess the proportion of viral nucleic acid present in each sample and determine the sequencing limits for whole genome retrieval using both the laboratory based Illumina technology and the portable MinION platform.

4.3 Results

4.3.1 Clinical samples viral load distribution

A total of 73 samples tested during 2016 in RIPL, PHE Porton Down, were positive by real-time qRT-PCR for CHIKV, and 368 were positive for DENV. Median Ct for CHIKV was 26.1, for DENV it was 26.8. To assess the feasibility of direct metagenomic sequencing of CHIKV and DENV genomes from patient serum/plasma, a set of positive samples were selected from samples collected during 2016 in the RIPL, PHE, Porton Down. For each virus, samples representing the range of viral titres seen during 2016 were selected, based on Ct value (Figure 4.1). CHIKV samples selected (n = 14) ranged from Ct 14.72 to Ct 32.57, corresponding to 10^{10} and 10^{5} genome copies per ml of plasma or serum. DENV samples selected (n = 12) ranged from Ct 16.29 to Ct 31.29, corresponding to 10^{9} and 10^{5} genome copies per ml (Table 4.1).

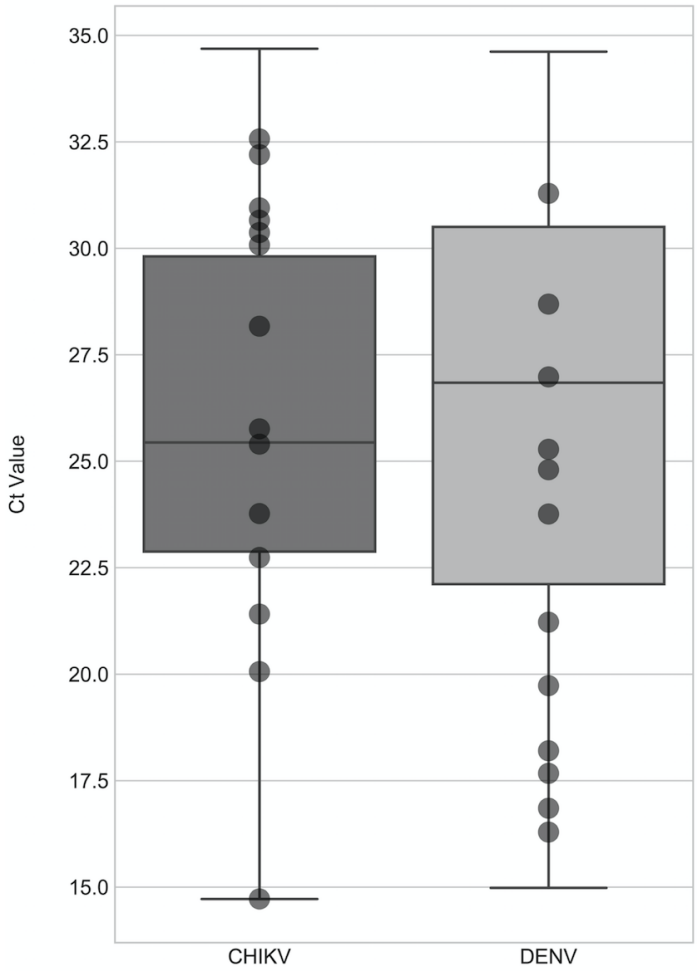


Figure 4.1 Cycle threshold values distribution of CHIKV (n = 73) and DENV (n = 368) positive samples from the Rare and Imported Pathogens Laboratory (n = 441 total samples).

The 14 CHIKV and 12 DENV samples selected for this work are indicated by circles. For each virus, the median Ct value of positive samples by quantitative real-time PCR is shown (horizontal line inside box), as well as, 25th and 75th percentiles (box lower and upper boundaries) and total range (whiskers).

4.3.2 Metagenomic MiSeq sequencing

To measure the proportion of viral nucleic acid present relative to host/background and assess achievable genome coverage, all samples were processed as described in methods and MiSeq sequenced (Table 4.1 and Table 4.2).

Table 4.1 Description of samples positive for CHIKV and DENV by real-time reverse transcription-PCR with corresponding MiSeq mapping data, (n = 26 samples)

Sample	Ct Value	Estimated copy number (/ml)	Sample type	Total reads (R1+R2)[a]	% Reads Mapping	% 20x coverage	% 10x coverage	Reference virus	Reference size (nts)
CHIKV 1	14.72	2.12E+10	Plasma	1113560	78.32%	99.59%	99.72%	CHIKV	11826
CHIKV 2	20.06	5.49E+08	Serum	1278624	98.48%	99.14%	99.47%	CHIKV	11826
CHIKV 3	21.41	2.18E+08	Plasma	1391258	95.23%	98.86%	99.37%	CHIKV	11826
CHIKV 4	22.74	8.76E+07	Plasma	888968	19.16%	97.08%	97.32%	CHIKV	11826
CHIKV 5	23.77	4.33E+07	Plasma	1357606	97.13%	99.16%	99.58%	CHIKV	11826
CHIKV 6	25.4	1.42E+07	Serum	3236848	34.88%	97.80%	98.40%	CHIKV	11826
CHIKV 7	25.76	1.11E+07	Plasma	3748070	72.77%	99.04%	99.56%	CHIKV	11826
CHIKV 8	28.17	2.13E+06	Plasma	1499952	28.41%	98.69%	99.00%	CHIKV	11826
CHIKV 9	30.08	5.76E+05	Serum	1035026	6.66%	95.98%	98.22%	CHIKV	11826
CHIKV 10	30.37	4.72E+05	Serum	1575222	16.84%	97.39%	98.01%	CHIKV	11826
CHIKV 11	30.66	3.87E+05	Serum	1143054	13.52%	95.36%	96.96%	CHIKV	11826
CHIKV 12	30.95	3.17E+05	Serum	1507380	10.93%	96.11%	96.52%	CHIKV	11826
CHIKV 13	32.2	1.35E+05	Serum	1323920	5.03%	88.47%	89.38%	CHIKV	11826
CHIKV 14	32.57	1.05E+05	Serum	1479404	21.72%	96.32%	96.93%	CHIKV	11826

Ct: cycle threshold;

[a] 'R1+R2' indicates paired-end sequencing

Table 4.2 Description of samples positive for CHIKV and DENV by real-time reverse transcription-PCR with corresponding MiSeq mapping data, (n = 26 samples)

Sample	Ct Value	Estimated copy number (/ml)	Sample type	Total reads (R1+R2)[a]	% Reads Mapping	% 20x coverage	% 10x coverage	Reference virus[b]	Reference size (nts)
DENV 1	16.29	4.21E+09	Plasma	439292	93.44%	99.51%	99.58%	DENV 1	10735
DENV 2	16.85	2.83E+09	Serum	513472	92.56%	99.40%	99.58%	DENV 1	10735
DENV 3	17.67	1.58E+09	Plasma	738814	92.53%	99.58%	99.58%	DENV 2	10723
DENV 4	18.20	1.09E+09	Serum	477368	93.97%	98.73%	99.12%	DENV 2	10723
DENV 5	19.73	3.67E+08	Serum	915554	89.65%	99.14%	99.40%	DENV 2	10723
DENV 6	21.22	3.61E+07	Serum	3587926	83.87%	99.68%	99.69%	DENV 4	10649
DENV 7	23.76	2.11E+07	Serum	4146678	2.17%	86.99%	89.13%	DENV 1	10735
DENV 8	24.8	1.01E+07	Serum	777264	69.23%	99.56%	99.58%	DENV 3	10707
DENV 9	25.28	7.17E+06	Plasma	787728	26.97%	98.77%	98.81%	DENV 2	10723
DENV 10	26.98	2.15E+06	Serum	596240	6.58%	93.47%	93.97%	DENV 3	10707
DENV 11	28.69	6.39E+05	Serum	1034698	3.73%	94.44%	94.70%	DENV 1	10735
DENV 12	31.29	1.01E+05	Serum	1374766	0.47%	71.46%	77.76%	DENV 1	10735

Ct: cycle threshold;

[a] 'R1+R2' indicates paired-end sequencing

[b] For DENV the serotype is also indicated

The proportion of total reads mapping to the respective viral reference was higher than anticipated for both viruses (Figure 4.2). In some low Ct samples, over 90% of reads mapped to the viral reference and proportions over 50% were still observed at mid-Ct range (Ct values between 20 and 25). The lowest proportion of reads mapping to the viral reference was 5.03% and 0.47% for CHIKV and DENV respectively (Table 4.1, Table 4.2, Figure 4.2). The majority of samples returned over 95% genome coverage at 20x (21/26 samples) and over 98% genome coverage at 10x (20/26 samples). Irrespective of lower mapping percentages in high Ct value samples, genome coverage of 88.5% (20x) and 89.4% (10x) for CHIKV and 75.0% (20x) and 77.8% (10x) for DENV was observed.

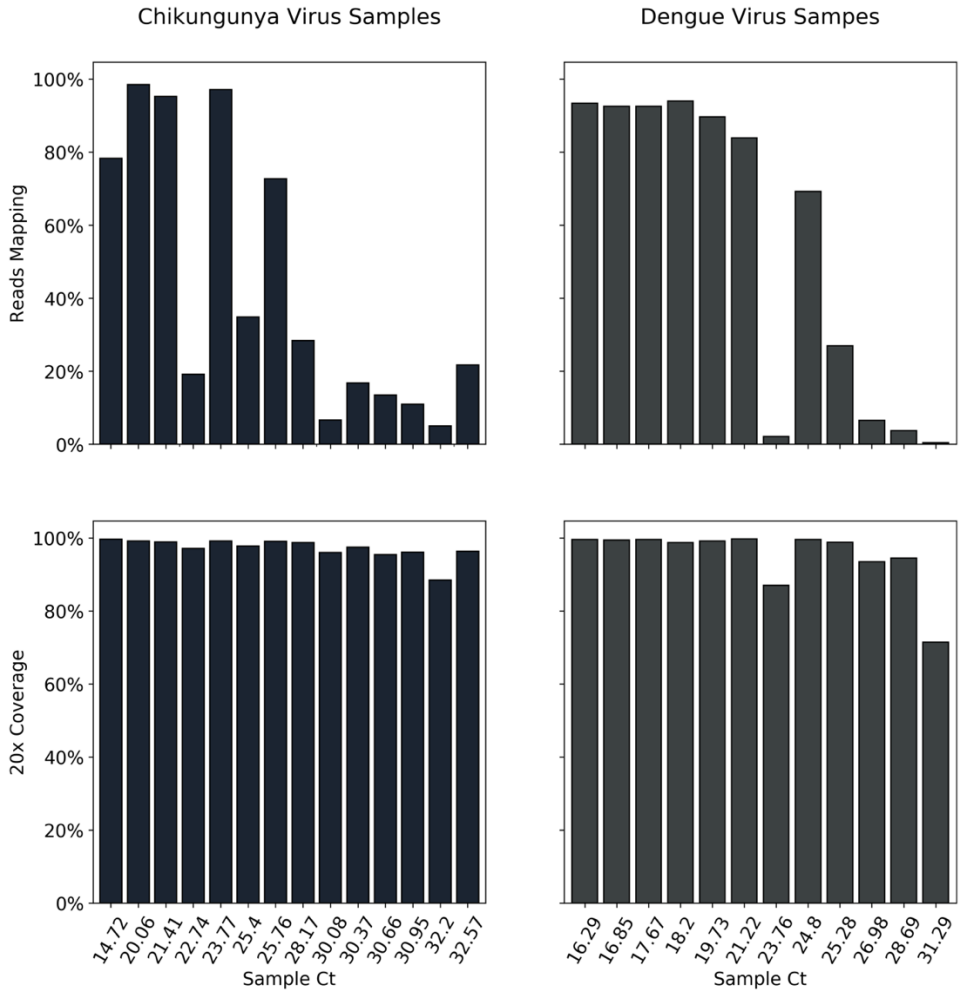


Figure 4.2 Proportion of reads mapping to the appropriate viral reference sequence and proportion of reference genome sequenced at minimum 20-fold coverage in each CHIKV or DENV positive sample (n = 26 samples).

The percentage of total reads mapping to the appropriate reference sequence is plotted in the upper panel. Lower panels display the percentage of the reference genome sequenced to a minimum depth of 20-fold in the MiSeq data.

A loose trend is observed between Ct value and proportion of viral reads, with a high level of per sample variation. The two lowest viral titre CHIKV samples (13 & 14) have similar Ct values (32.2 & 32.57) but vary significantly in the proportion of reads that are viral in origin (5.03% & 21.72%) (Figure 4.2, Table 4.1). The 5.03% viral reads in CHIKV13 is the lowest for CHIKV, yet still sufficient to generate 88.5% of the CHIKV genome at 20x depth from just ~662,000 paired-end reads. This amount of genomic information is highly informative and further sequencing would very likely increase coverage. Of all diagnostic samples tested in 2016 only 7 of the 73 CHIKV samples had a Ct greater than 32.2 (including sample CHIKV14) (Figure 4.1), which clearly indicates that for the great majority (>90%) of CHIKV PCR positive samples the viral load is sufficient for genome sequencing directly from patient samples without further enrichment beyond a simple DNAse digestion (Figure 4.1). For the DENV samples, the lowest viral read proportion observed was 0.47% in DENV 12, Ct 31.29 (Figure 4.2, Table 4.2). This generated 71.5% coverage at 20x depth (increased to 77.8 at 10x depth) from just 687,000 paired end MiSeq reads and allowed for DENV serotype identification. Only 62 of the 368 DENV cases in 2016 had a higher Ct, predicting that >80% of PCR positive DENV samples have a viral load sufficient for genome sequencing (Figure 4.1).

4.3.3 Metagenomic MinION sequencing

Four representative samples for each virus were selected for MinION sequencing. Samples were selected with low, mid and high Ct values. CHIKV 1, CHIKV 3, CHIKV 4 and CHIKV 9 (Ct values: 14.72, 21.41, 22.74 and 30.08) and DENV 1, DENV 2, DENV 6, DENV 11 (Ct values: 16.29, 16.85, 21.22 and 28.69). Total cDNA from the SISPA preparation was used as input for the sequencing library preparation and sequenced individually on a single flow cell. Details on total input amounts, sequencing kit and flow cell information for each sample can be found in Table 4.3. Total 1D reads generated for each run varied between 203,000 and 3,481,000 per sample and mean read length ranged from 564 to 886 bases (Table 4.3).

Table 4.3 Description of CHIKV and DENV positive samples by real-time reverse transcription-PCR and corresponding MinION sequencing data (n = 8 samples).

Sample	Ct Value	Input (ng)	Sequencing Kit (2D kit version)	Flow cell (FLO-)	1D Total bp	1D Total Reads	1D Mean Read Length (nt)	1D Max Read Length (nt)
CHIKV 1	14.7	431.5	SQK-NSK007	MIN105	1.51E+08	267171	564	92712
CHIKV 3	21.4	928.8	SQK-LSK208	MIN106	1.63E+09	1891028	862	99031
CHIKV 4	22.7	113.4	SQK-NSK007	MIN105	1.74E+08	216493	805	125387
CHIKV 9	30.1	212.4	SQK-LSK208	MIN106	2.12E+09	3481358	608	121711
DENV 1	16.3	1626.0	SQK-NSK007	MIN105	2.42E+08	284622	851	115494
DENV 2	16.9	1626.0	SQK-NSK007	MIN105	1.55E+08	203700	760	52157
DENV 6	21.2	475.0	SQK-LSK208	MIN106	1.22E+09	1377721	886	118733
DENV 11	28.7	65.8	SQK-LSK208	MIN106	7.07E+08	1111566	636	119438

Figure 4.3 shows the percentages of reads mapping to viral reference which were generally concordant with the MiSeq data, although a slight decrease is observed across the range of Ct values. In the MinION data the highest mapped read percentages observed were 85.12% and 72.14% for CHIKV 9 and DENV 2 respectively, compared to 95.23% and 92.56% in the MiSeq data from the same samples. Whilst in high Ct samples the viral proportion drops to 4.08% for CHIKV 9 and 2.90% for DENV 11, from 6.66% and 3.73% in the MiSeq data. Despite the decrease in proportion of mapped viral reads, comparable genome coverage is observed at both 20x and 10x (Figure 4.3, Table 4.4) and is even increased compared to MiSeq data at lower viral titres, e.g. 100% at 20x for CHIKV 9 compared to 95.98% in the MiSeq data and 95.25% for the high Ct DENV 11 sample which generated 94.44% coverage from the MiSeq data. Differences in precise proportions of viral reads seen are likely due to inter-library variation. Differences in genome coverage achieved are due to both differences in total reads generated per sample (not normalised between platforms) as well as differences in average read length. Average read lengths in MinION data ranged from 564 to 886 bp (Table 4.3).

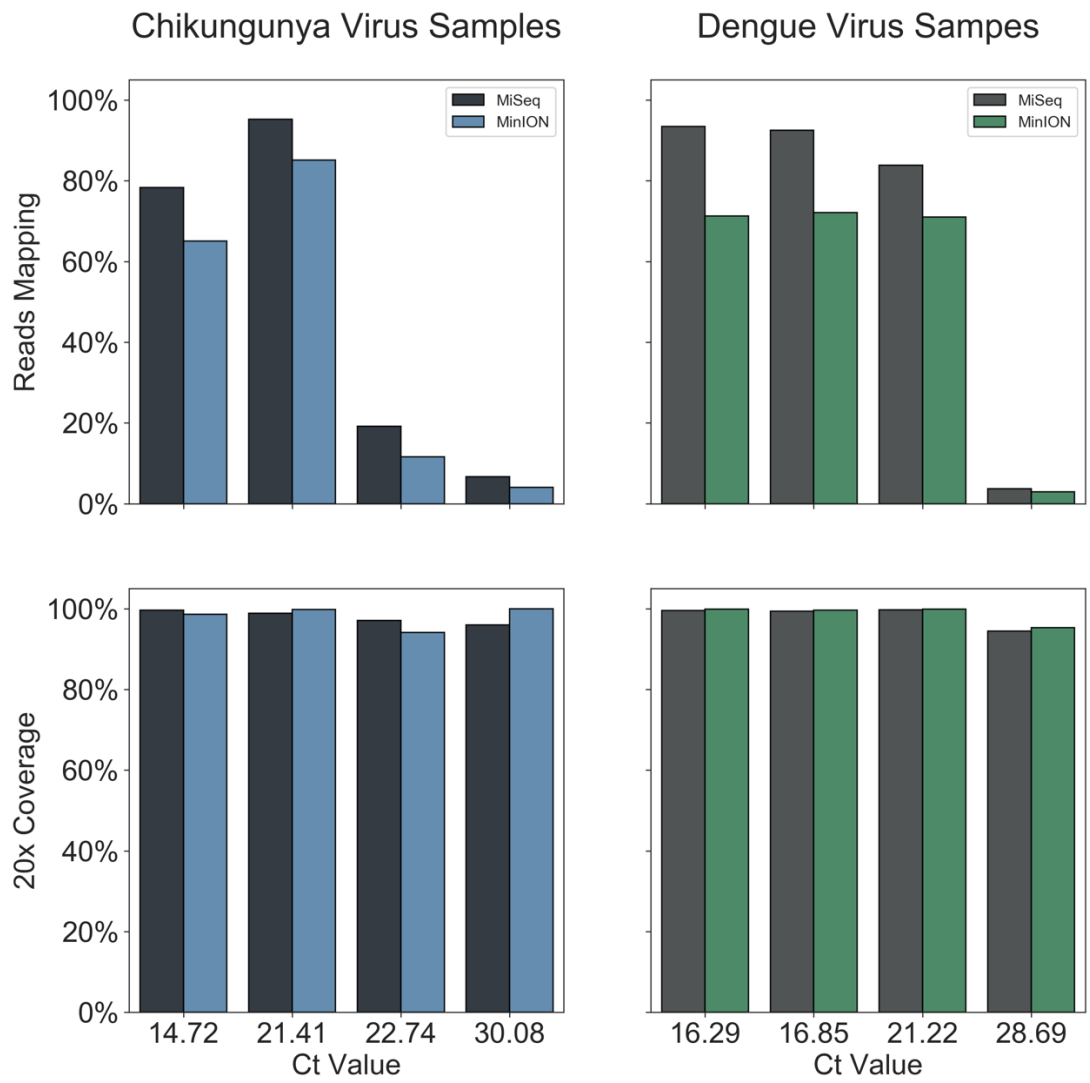


Figure 4.3 Comparison of MinION and MiSeq results, as to proportions of reads mapping to the appropriate reference viral sequence, and proportions of reference genome sequenced at minimum 20-fold coverage (n = 8 samples).

The percentage of total reads mapping to the appropriate reference sequence is plotted in the upper panel. Lower panels display the percentage of the reference genome sequenced to a minimum depth of 20-fold in the data generated, in dark blue or dark green for the MiSeq sequence data, in light blue or light green for MinION data (MinION_TC).

Table 4.4 Summary of MinION mapping data on CHIKV and DENV positive samples (n = 8 samples)

Sample	Ct Value	Total Reads	% Reads Mapping	% 20x Coverage	20x Genome Length (nt)	% 10x Coverage	Reference	Reference Size (nt)	Max *de Novo* Contig (nt)
CHIKV 1	14.7	267171	65.1%	98.57%	11658	99.2%	CHIKV	11826	5263
CHIKV 3	21.4	1891028	85.1%	99.76%	11798	99.9%	CHIKV	11826	10793
CHIKV 4	22.7	216493	11.6%	94.11%	11130	97.2%	CHIKV	11826	4256
CHIKV 9	30.08	3481358	4.08%	100%	11826	100%	CHIKV	11826	9860
DENV 1	16.3	284622	71.3%	99.9%	10719	99.9%	DENV 1	10735	8281
DENV 2	16.9	203700	72.1%	99.6%	10692	99.6%	DENV 1	10735	10157
DENV 6	21.2	1377721	71.1%	99.9%	10634	99.9%	DENV 4	10649	7877
DENV 11	28.7	1111566	2.9%	95.3%	10226	96.3%	DENV 1	10735	4699

Figure 4.4 shows coverage depth of reads mapped across the relevant genome for each sample sequenced by both MiSeq and MinION. Read levels are not normalised thus actual depth is a function of total reads sequenced, but the pattern of coverage seen is highly similar suggesting it is more dependent upon the SISPA methodology than sequencing library preparation. From MinION consensus genome sequences, between 99.93% and 100% of bases called per sample agreed with the MiSeq generated sequence.

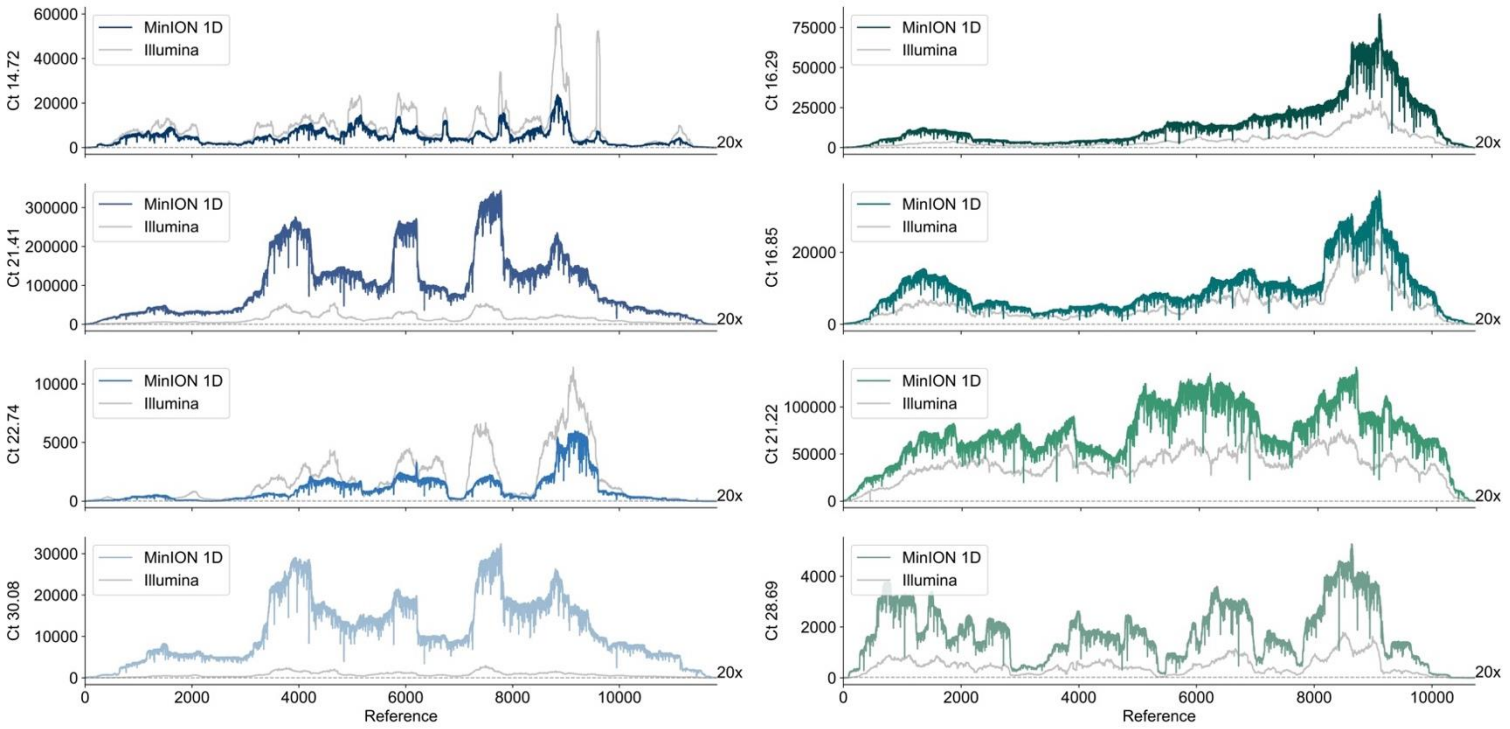


Figure 4.4 Coverage depth across the CHIKV or DENV genome, (n = 8 samples).

Each graph corresponds to a given sample, defined by its Ct value. Read depth (y-axis) across the genome (x-axis) following reference alignment is shown. MiSeq coverage is shown in darker blue and darker green for CHIKV and DENV positive samples respectively. MinION coverage is indicated in lighter blue or lighter green for CHIKV and DENV positive samples respectively. Total depth has not been normalised; comparison is to show overall pattern of coverage is highly similar across the methods. Dotted horizontal line indicates depth of 20x coverage, used for consensus calling.

4.3.4 Metagenomic data analysis and co-infection identification

To test the applicability of a metagenomic analysis approach to the data, we assessed read taxonomic classification using Kraken (Figure 4.5). The distribution of reads classified as CHIKV, DENV, other viruses, bacteria, and archaea/eukaryota show a similar pattern for MiSeq and MinION data. The proportion of unclassified reads for each sample increased with Ct value, as the proportion of human origin reads is higher and the human genome is not represented in the Kraken database used. A decrease in the percentage of CHIKV and DENV classified reads is observed for MinION data compared to MiSeq, but was sufficient to identify the correct predominant virus in all samples.

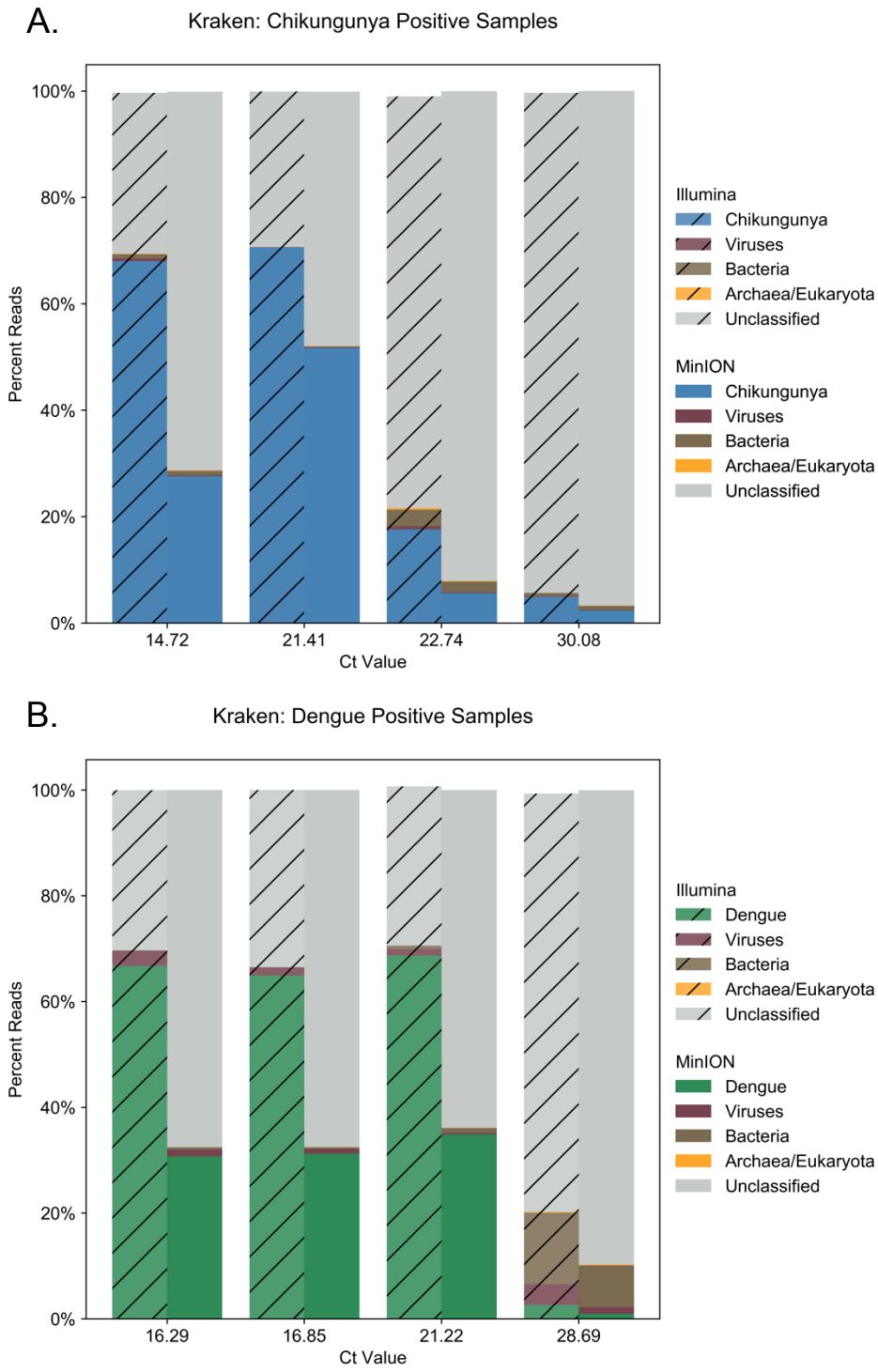


Figure 4.5 Kraken classification of reads from metagenomic sequencing in (A) CHIKV and (B) DENV real-time reverse transcription-PCR positive samples (n = 8 samples). Kraken classification distribution comparison for MiSeq (cross-hatched) and MinION data. Reads grouped as either CHIKV (blue in panel A), DENV (green in panel B), other viruses (brown), archea/eukaryota (orange), bacteria (brown) or unclassified (grey).

Kraken analysis also allowed for the identification of a DENV co-infection in sample CHIKV 3, the consensus sequence of which was unique in the sample set,

eliminating cross-contamination from the DENV positive samples as potential source. Kraken classified 0.08% of MiSeq reads and 0.15% of MinION reads as DENV. Using reference mapping to validate the finding, 0.22% of MiSeq reads and 0.43% of MinION reads mapped to a DENV-1 reference genome. Genome coverage at 20x of 99.73% and 95.99% was achieved for the primary CHIKV and secondary DENV co-infection respectively, with a single MinION flow cell.

4.3.5 *De novo* assembly

An alternative reference-free approach to read classification *de novo* assembly of the data was attempted using Canu (*338*) and contig identification using Basic Local Alignment Search Tool against the GenBank nt database (BLASTn). Table 4.4 lists the longest viral contig length identified in each sample, ranging from 4.2 Kb (36% of reference genome size) to 10.8 Kb (91%) for CHIKV and 4.7 Kb (44%) to 10.1 Kb (95%) for DENV. Identification of the pathogen present without prior knowledge would have therefore been possible for all samples.

4.3.6 Updated MinION library kits

In order to keep up with the rapidly evolving MinION chemistries and flowcell developments we repeated the sequencing of the co-infected CHIKV 3 sample utilising the MinION $1D^2$ (SQK-LSK308) and Rapid (SQK-RBK001) kits, currently the most accurate and the fastest library preparation kits available, respectively. Using the $1D^2$ kit 74.5% of reads generated mapped to CHIKV and 0.37% to DENV, while from the Rapid kit the result was 66.26% and 0.29% respectively (both lower than observed in the 2D chemistry). Coverage at 20x for CHIKV was above 99% for both kits, and for DENV was 95.04% from the $1D^2$ and 81.09% from the Rapid kit (Table 4.5). Coverage depth pattern across the genome for both viruses (Figure 4.6) was similar for all library kits tested. Near-maximum coverage for both viruses was obtained within 2000 sec (~30 minutes) with the 2D kit, 500 sec (~8 minutes) with the $1D^2$ kit and 5000 sec (~85 minutes) with the Rapid kit (Figure 4.7). *De novo* assembly (Table 4.5) produced best CHIKV contigs of 10.7, 11.3 and 11.4 Kb for the 2D, $1D^2$ and Rapid libraries respectively and the longest contigs generated for DENV were 7.5, 2.3 and 4.2 Kb.

Table 4.5 Comparison of MinION mapping data across library kits (n = 8 samples)

Platform	Kit Information	Flowcell (FLO-)	Virus Identified	Total Reads (nt)	% Reads Mapping	% 20x Coverage	% 10x Coverage	Reference	Reference Size (nt)	Max *de Novo* Contig (nt)
Illumina	Nextera XT	NA	CHIKV	1391258	95.23%	98.86%	99.37%	CHIKV	11826	7321
Illumina	Nextera XT	NA	DENV	1391258	0.22%	63.66%	77.82%	DENV1	10735	6613
MinION 2D	SQK-LSK208	MIN106	CHIKV	1891028	85.12%	99.73%	99.91%	CHIKV	11826	10793
MinION 2D	SQK-LSK208	MIN106	DENV	1891028	0.43%	95.99%	96.09%	DENV1	10735	7549
MinION $1D^2$	SQK-LSK308	MIN107	CHIKV	5080906	74.50%	99.94%	100%	CHIKV	11826	11369
MinION $1D^2$	SQK-LSK308	MIN107	DENV	5080906	0.37%	95.04%	96.42%	DENV1	10735	2199
MinION Rapid	SQK-RBK001	MIN106	CHIKV	611110	66.26%	99.66%	99.68%	CHIKV	11826	11473
MinION Rapid	SQK-RBK001	MIN106	DENV	611110	0.29%	81.09%	90.83%	DENV1	10735	4227

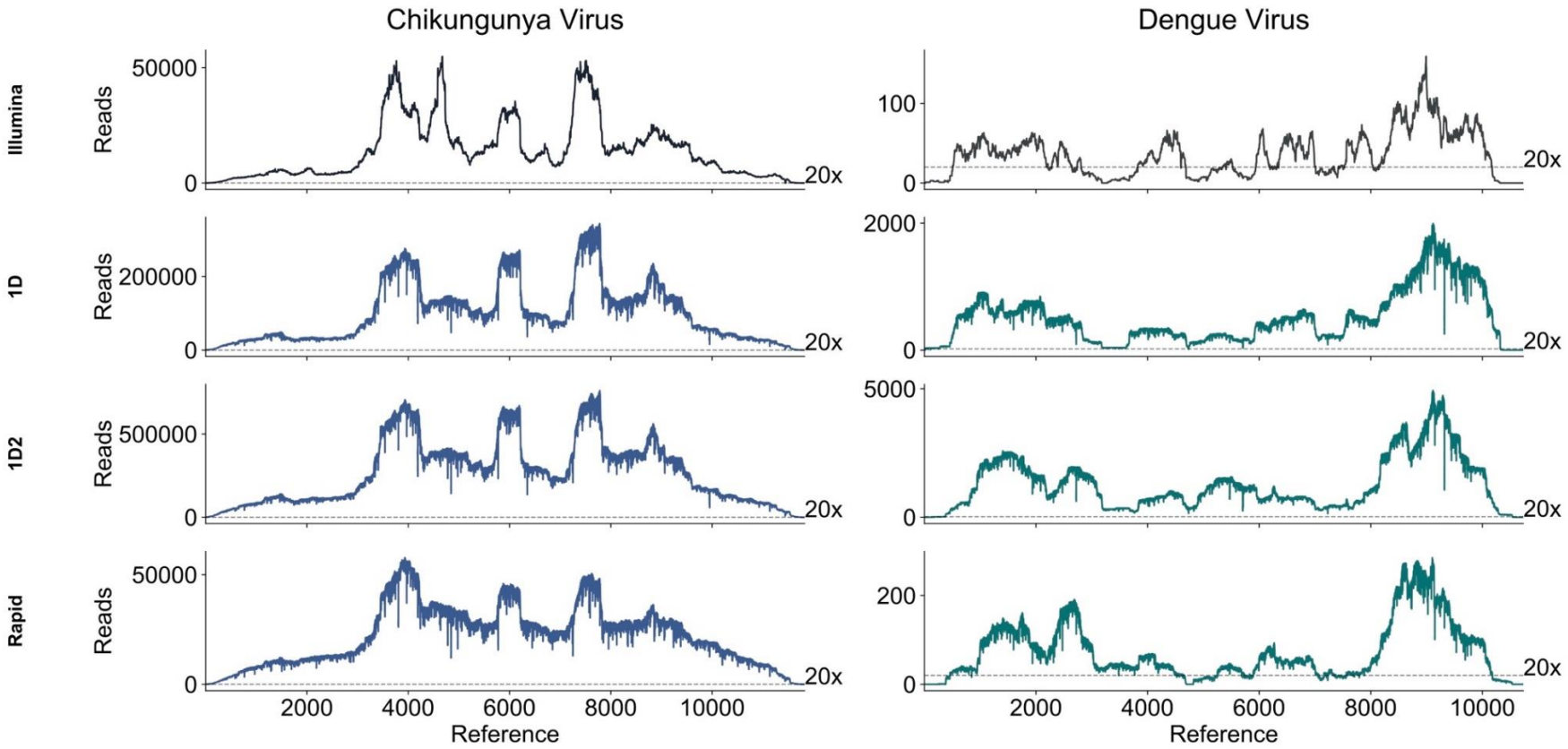


Figure 4.6 Comparison of genome coverage depth across the CHIKV virus or DENV genome for different sequencing library preparation methods in a sample coinfected with DENV and CHIKV viruses (n = 1 sample).

Read depth across both CHIKV and DENV genomes following reference alignment is shown for coinfection sample CHIKV 3, sequenced using four different sequencing library preparation/sequencing methods. Total coverage depth has not been normalised; comparison is to show overall pattern of coverage is highly similar across the methods. Dotted horizontal line indicates depth of 20x coverage, used for consensus calling.

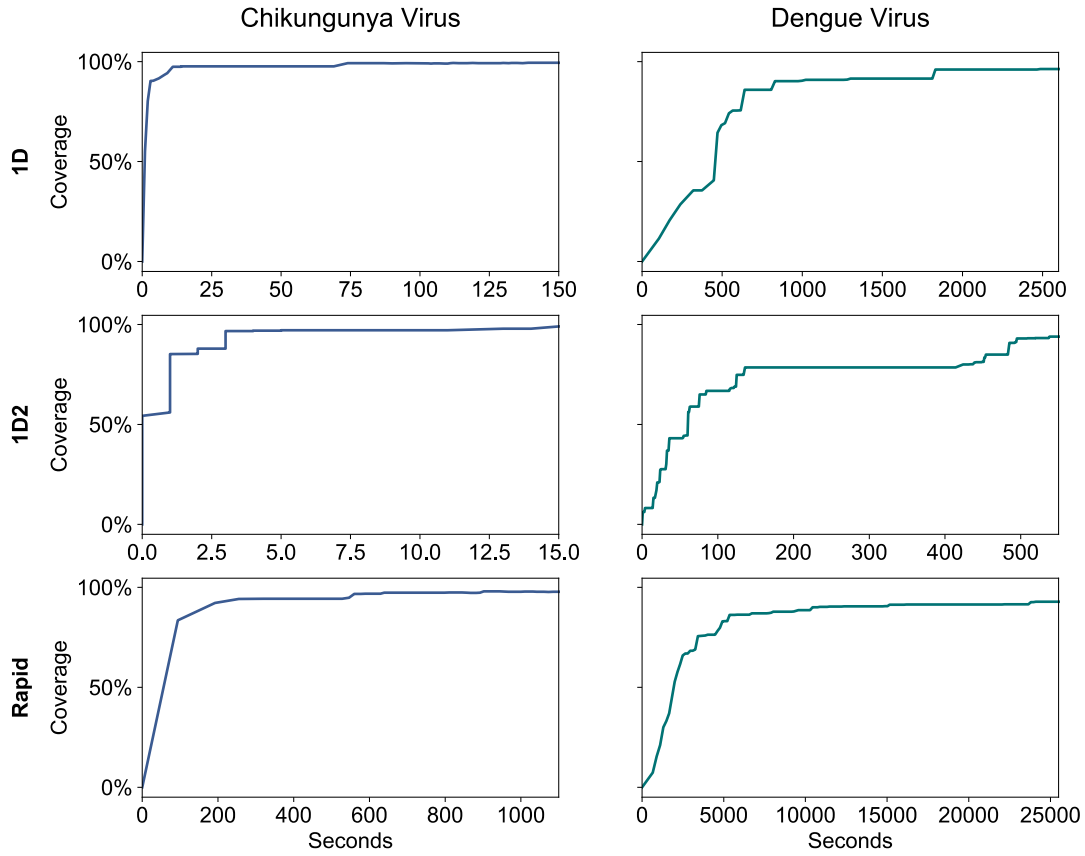


Figure 4.7 Proportion of genome covered over the course of each sequencing run. The percentage of the CHIKV or DENV genome sequenced plotted over the course of the MinION sequencing run for each kit version tested.

The 1D data from the Rapid kit was sufficient to call a consensus from 11647/11826 bases of the CHIKV reference with 179/11826 bases called as ambiguous or too low coverage. All bases called were concordant with the MiSeq consensus. A polishing step using Nanopolish (*320*)(*320*) with a subset of the mapped reads (~100x coverage depth) significantly reduced ambiguous calls to 90/11826, introducing a single disagreement with the MiSeq consensus (99.99% concordance). Despite considerably greater read depth the $1D^2$ kit called only 11082/11826 due to a higher

proportion, 744/11826, of ambiguous base calls. Suggesting 1D reads are most suitable for this approach.

4.4 Discussion

These results clearly show that there are considerable levels of viral nucleic acid present in a large proportion of CHIKV and DENV real-time qRT-PCR positive clinical samples, and demonstrate that relatively modest metagenomic sequencing is capable of elucidating significant portions of viral genome even for samples with Ct values at the higher end of clinical range. A decreased Ct value coincided with an increased proportion of viral reads, with a considerable level of variation between samples, likely because of the total level of non-viral host/background nucleic acid present due to variability between patients or in sample handling during collection, storage and testing. For example, the two lowest viral titre CHIKV samples (13 and 14) have similar Ct values (32.2 and 32.57) but varied significantly in the proportion of viral reads (5.03% and 21.72%). The 5.03% viral reads in CHIKV13 is the lowest for CHIKV, yet still sufficient to generate 88.5% of the CHIKV genome at 20x depth from just ca 662,000 paired-end MiSeq reads. This amount of genomic information is highly informative and further sequencing would likely increase coverage. Only seven of the 73 total CHIKV diagnostic samples tested in 2016 had a Ct greater than 32.2 (including sample CHIKV14) (Table 4.1), which suggests that for the majority (> 90%) of CHIKV PCR positive samples, viral load is sufficient for genome sequencing directly from patient samples without further viral enrichment beyond a simple DNAse digestion (Figure 4.4). The lowest viral read proportion observed in the DENV samples was 0.47% in DENV12, Ct 31.29, which generated 71.5% coverage at 20x depth (increased to 77.8 at 10x depth) from just 687,000 paired end MiSeq reads and allowed for DENV serotype identification. Only 62 of 368 DENV cases in 2016 had a higher Ct, predicting that > 80% of PCR positive DENV samples have a viral load sufficient for genome sequencing (Figure 4.4). These estimates are based on Ct range distribution from a single year, results may vary from year to year.

The high yield of viral sequences from clinical CHIKV and DENV samples make the exciting prospect of metagenomic MinION viral whole-genome-sequencing feasible, even for lower viral titre samples. Evaluating this on a representative subset of our samples demonstrates that viral read proportions are in general agreement with that seen for MiSeq sequencing, predicting a similar proportion of real-time qRT-PCR positive patient samples would be suitable for direct metagenomic sequencing on the MinION. Differences in precise proportions of viral reads seen between MiSeq

and MinION are likely due to inter-library variation. Differences in genome coverage achieved are due to both differences in total reads generated per sample (not normalised between platforms) as well as differences in average read length. Of the samples tested on the MinION, the lowest titre samples CHIKV 9 and DENV 11 both generated near complete genome coverage.

We repeated the sequencing of the coinfected CHIKV 3 sample using the MinION $1D^2$ (SQK-LSK308) and Rapid (SQK-RBK001) kits. A reduction in viral proportion of total reads was observed compared with the 2D kit, which may be due partly to the extended storage time of the original samples before retesting. In the case of the $1D^2$ kit, the lower proportion was outweighed by a substantial increase in total data generated per flow cell (5 M vs 1.8 M reads). For the Rapid kit, the total data produced should be considered in the light of the greatly simplified sample workflow and turnaround-time.

The use of metagenomics to elucidate genomic sequences of RNA viruses directly from clinical samples has several obvious benefits in public health applications. The primary benefit over targeted methods is the hypothesis-free nature of the assay, which allows identification and genomic characterisation of novel or unexpected RNA viral agents, either as primary or coinfectants (demonstrated here in the CHIKV/DENV coinfection sample), without any prior clinical knowledge. It also removes the need for laboratory optimisation of targeted methods, such as primer or bait-probe design and testing, and is not subject to escape mutations in target sites that afflict targeted sequencing and diagnostic methods. This issue is particularly relevant for highly diverse RNA viruses, such as LASV, which are difficult to assess using targeted methods, without regular reappraisal (*48*).

The principal limitation of the metagenomic approach is the limit of detection. The data generated here show that viral titres as low as 10^5 are sufficient for significant genome recovery by this method, but ZIKV is a recent example of a pathogen typically present at lower clinical titres, for which targeted methods are an absolute requirement (*156*, *218*). For diagnostic purposes real-time qRT-PCR has a lower limit of detection, provided the target site is conserved in the pathogen isolate tested. Clearly no single method is most suitable for both detection and genotyping of all pathogens and each has a role to play in differing circumstances.

4.5 Conclusions

These results demonstrate that across the clinically relevant range of viral loads an unexpectedly high proportion of reads generated metagenomically from CHIKV and DENV clinical samples are viral in origin. Direct metagenomic sequencing of nucleic acid extracts from serum and plasma without viral enrichment allowed for virus and coinfection identification, subtype determination and in the majority of cases elucidated complete or near-complete genomes adequate for phylogenetic analysis. Therefore, metagenomic sequencing provides an effective approach for the analysis of CHIKV and DENV genomes directly from the majority of real-time qRT-PCR positive serum and plasma samples, without the need for culture or viral nucleic acid enrichment beyond a simple DNA digestion. The results presented in this chapter along with recent improvements in sequencing library preparation and total read numbers generated by the MinION, make metagenomic sequencing uniquely valuable for whole genome sequencing of these and likely other RNA viruses; particularly valuable due to its portability, enabling diagnostic and outbreak support in remote and resource limited settings.

Chapter 5

Metagenomic sequencing at the epicentre of the Nigeria 2018 Lassa fever outbreak

5. Chapter 5. Metagenomic sequencing at the epicentre of the Nigeria 2018 Lassa fever outbreak

5.1 Overview

To evaluate metagenomic nanopore sequencing for the recovery of whole viral genome sequences from clinical samples of a divergent virus in a remote and resource-limited setting, metagenomic nanopore sequencing of LASV was implemented at ISTH, Edo State, Nigeria during the 2018 season. The original pilot-scale study was expedited as the 2018 Nigerian Lassa fever season saw the largest ever recorded upsurge of cases, raising concerns over the emergence of a strain with increased transmission rate. A total of 120 samples were sequenced in-country and on-site computational analysis facilitated the generation of 36 near-complete LASV genomes. The real-time analysis of the 36 genomes and the subsequent confirmation using all 120 samples sequenced in the country of origin revealed extensive diversity and phylogenetic intermingling with strains from previous years, suggesting independent zoonotic transmission events and thus allaying concerns of an emergent strain or extensive human-to-human transmission.

5.2 Introduction

Portable field methodologies are vital for pathogen monitoring, especially when epidemics and outbreaks occur in resource-limited settings. Real-time field nanopore sequencing utilising PCR-amplicon based approaches provided important data during the 2014-2016 EBOV outbreak in West Africa (*147*) the 2015-2016 ZIKV epidemic (*218*) and the 2016-2017 YFV outbreak in Brazil (*156*, *219*). However such an approach is extremely challenging for highly variable pathogens such as LASV, for which even PCR-based laboratory diagnosis poses a significant challenge, due to its high inter-strain nucleic acid sequence variation (*48*). Designing targeted whole-genome sequencing approaches, such as PCR amplicons or bait/capture probes, without prior knowledge of the specific LASV lineage to target is therefore impractical. Chapters 3 and 4 highlighted the feasibility of retrieving complete viral genomes directly from patient samples at clinically relevant viral titres using this approach for DENV and CHIKV (*339*).

A pilot study was designed to establish and evaluate field metagenomic sequencing using the MinION in a resource limited setting during a LASV season in

an endemic region of Nigeria. The aim was to test and troubleshoot metagenomic MinION sequencing on-site, alongside conventional diagnostic capacity and to generate complete consensus sequences for this highly diverse pathogen. Our pilot study was expedited, due to a large upsurge in the number of cases compared to previous years, which raised concerns over the emergence of a strain with increased transmission rate. This chapter describes the application of metagenomic sequencing for LASV in a resource-limited setting during a seven-week deployment at ISTH, Edo State, concurrent with the 2018 LASV season. A total of 120 LASV positive samples were sequenced during the seven week deployment, selected from the 376 cases reported by ISTH between 1st January and 16th March 2018. To reflect the geographic case distribution of the outbreak, the majority of samples originated from Edo state followed by Ondo and Ebonyi. Samples selected covered the wide range of clinical viral loads observed, including several samples testing negative in one of the two real-time qRT-PCR assays used. To understand the molecular epidemiology of this upsurge metagenomic nanopore sequencing was performed directly on this representative set of patient samples and genomic data along with phylogenetic reconstructions were communicated immediately to Nigerian authorities and the WHO to inform the public health response.

5.3 Results

5.3.1 Metagenomic MiSeq Sequencing

Positive LASV samples were selected from a cohort of samples collected during the Nigerian endemic seasons of LASV in 2014, 2016 and 2017. Samples (n = 15) were selected to represent the range of Ct values observed during the endemic seasons, based on the Altona real-time qRT-PCR assay Ct values and ranged from Ct 14.55 to Ct 33.44 (Table 5.1). To measure the proportion of viral nucleic acid present relative to host/background and assess the genome coverage, all samples were processed as described in Section 2.8 and MiSeq sequenced as described in Section 2.9.

Table 5.1 Description of samples positive for LASV by Altona real-time reverse transcription-PCR and Ct values for the ones also tested by Nikisins real-time reverse transcription-PCR (if none is stated information is not available) with corresponding MiSeq reads (n = 15 samples)

Sample ID	Year of sample collection	Altona Ct value	Nikisins Ct value	Sample type	Total reads (R1+R2)[a]
LASV01	2016	14.55	None	Serum	480136
LASV02	2017	16.75	18.34	Serum	726608
LASV03	2016	16.92	None	Serum	648097
LASV04	2014	18.66	None	Serum	641387
LASV05	2017	19.42	19.25	Serum	559097
LASV06	2017	21.01	21.22	Serum	758812
LASV07	2017	23.07	None	CSF	850114
LASV08	2017	24.50	None	Serum	487299
LASV09	2017	25.66	None	Serum	567651
LASV10	2017	26.07	None	Serum	868797
LASV11	2017	28.97	19.02	Serum	790793
LASV12	2017	30.71	None	Serum	724618
LASV13	2017	31.82	None	CSF	446391
LASV14	2017	32.32	None	Serum	633164
LASV15	2017	33.44	None	Serum	839807

[a] 'R1+R2' indicates paired-end sequencing

Due to the diversity of LASV a *de novo* approach was used to assemble the data and the generated contigs were entered in a blastn search to identify the most suitable reference for read mapping to generate consensus. Six samples assembled a substantial proportion of the L Segment (>50%) and 5 generated significant contigs (>14% of the L Segment). For the majority of the samples (n = 11) near complete (> 75%) S Segments were *de novo* assembled. Contigs generated were sufficient for the identification of L segment reference sequences for 14 samples and for the identification of S segment reference sequences for all 15 samples (Table 5.2).

Table 5.2 Summary of LASV positive sample de novo assemblies and reference identification for each segment

Sample ID	Altona Ct value	Max de novo contig L segment	Max de novo contig S segment	L segment reference	S Segment reference
LASV01	14.55	6942	3591	KM822052.1	KM822014.1
LASV02	16.75	3196	3329	LC387475.1	GU481068.1
LASV03	16.92	6066	3417	KM822087.1	KM822088.1
LASV04	18.66	5726	3496	KM821956.1	KM821957.1
LASV05	19.42	1066	411	KT992433.1	MK107930.1
LASV06	21.01	3090	2945	MH053570.1	MH053573.1
LASV07	23.07	7321	3458	KM822048.1	KM822049.1
LASV08	24.50	647	2630	MH053547.1	MH053575.1
LASV09	25.66	279	1247	MH053479.1	MH053468.1
LASV10	26.07	1482	3409	MH053547.1	MH053586.1
LASV11	28.97	6982	3447	MH053469.1	MH053472.1
LASV12	30.71	4592	2832	KM822061.1	KM822089.1
LASV13	31.82	2337	2758	MH887950.1	KM822101.1
LASV14	32.32	410	860	KM822083.1	GU481076.1
LASV15	33.44	-	815	NA	MH053528.1

The proportion of total reads mapping to the respective reference was low across the whole range of Ct values (Table 5.3, Figure 5.1). For both segments combined the highest proportion (41.18%) was observed for sample LASV01 with Ct value 14.55, followed by sample LASV11 (Ct value: 28.97, reads mapping: 12.69%), LASV07 (Ct value: 23.07, reads mapping: 6.27%) and LASV04 (Ct value: 18.66, reads mapping: 4.10%). Figure 5.1 shows the percentages of reads mapping to the respective viral reference for each segment separately along with the 20x percentage genome coverage succeeded for each. Irrespective of the relatively low amount of total reads mapping to LASV, genome coverage (20x) of >80% was observed for 7 samples for the L segment and 9 samples for the S segment. A genome recovery of >50% was successful for 9 samples for the L segment and 12 samples for the S segment.

Table 5.3 Description of LASV positive samples by real-time reverse transcription-PCR and corresponding read mapping percentages (n = 15 samples)

Sample ID	Altona Ct Value	Nikisins Ct Value	% reads mapping	L segment % reads mapping	L segment % 20x coverage	S segment % reads mapping	S segment % 20x coverage
LASV01	14.55	None	41.18%	20.90%	100.00%	20.28%	100.00%
LASV02	16.75	18.34	1.56%	0.73%	89.87%	0.83%	99.20%
LASV03	16.92	None	1.19%	0.73%	94.70%	0.46%	96.14%
LASV04	18.66	None	4.10%	2.74%	97.89%	1.36%	97.51%
LASV05	19.42	19.25	0.38%	0.24%	36.64%	0.14%	60.71%
LASV06	21.01	21.22	2.13%	0.71%	86.15%	1.42%	98.06%
LASV07	23.07	None	6.27%	3.12%	100.00%	3.15%	99.59%
LASV08	24.50	None	0.23%	0.10%	12.37%	0.13%	67.88%
LASV09	25.66	None	0.24%	0.03%	3.66%	0.21%	17.19%
LASV10	26.07	None	0.20%	0.06%	20.35%	0.14%	84.85%
LASV11	28.97	19.02	12.69%	6.56%	99.73%	6.13%	99.94%
LASV12	30.71	None	0.20%	0.12%	58.99%	0.08%	68.26%
LASV13	31.82	None	0.59%	0.36%	60.29%	0.23%	84.43%
LASV14	32.32	None	0.44%	0.05%	5.74%	0.39%	26.41%
LASV15	33.44	None	0.01%	0.00%	0.00%	0.01%	4.58%

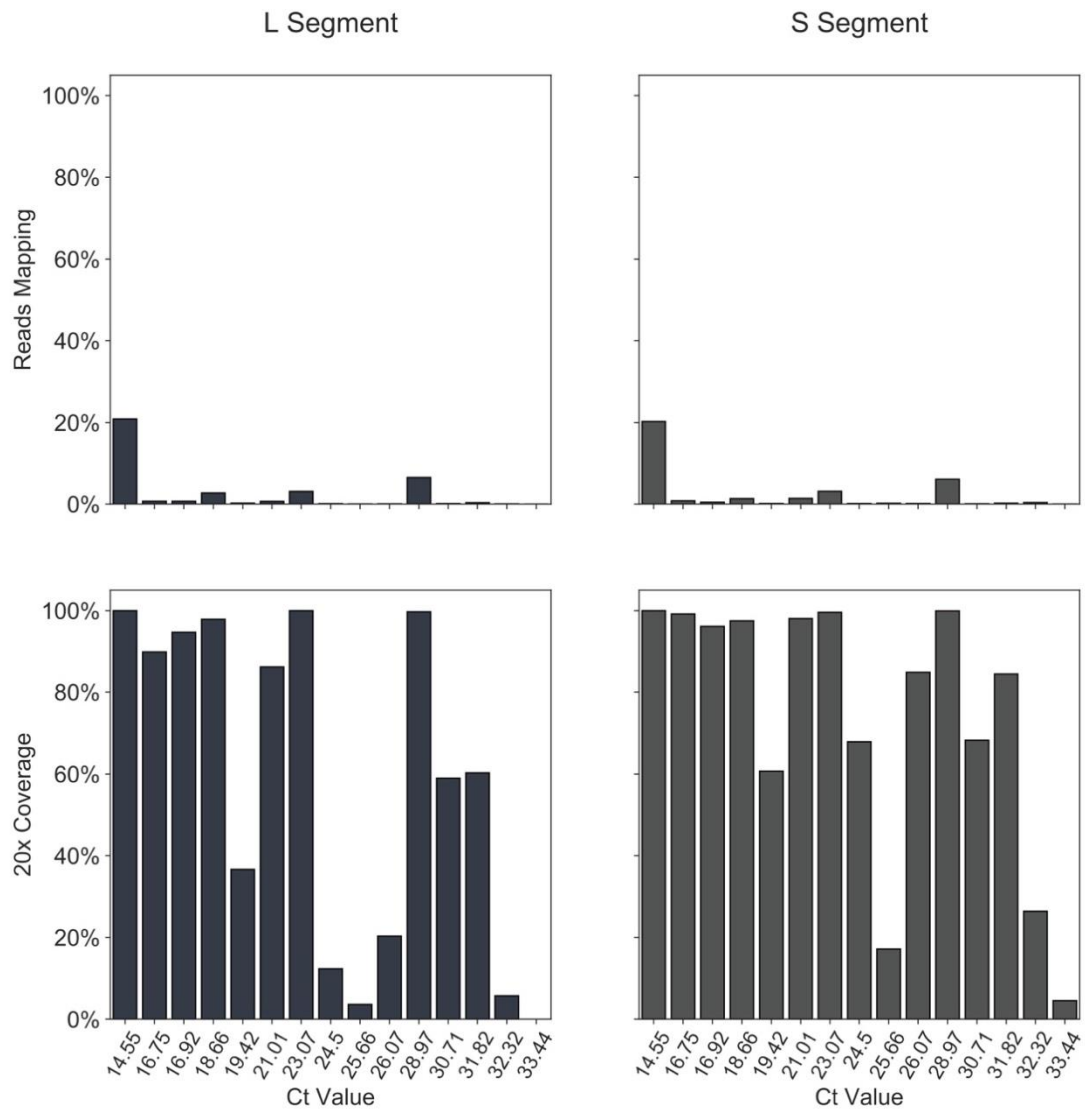


Figure 5.1 Proportion of reads mapping to the appropriate viral reference for each segment separately and proportion of reference genome recovered at minimum 20-fold coverage for each segment (n = 15 samples).

The percentage of total reads mapping to each segment reference is plotted in the upper panel. Lower panels display the percentage of the segment reference genome sequenced to a minimum depth of 20-fold in the MiSeq data.

5.3.2 Platform comparison and nanopore method validation

A subset of LASV samples were selected from the cohort of samples identified positive during the 2018 Nigerian outbreak. A total of 14 samples were randomly selected for sequencing on both platforms, with samples covering a range of Altona real-time qRT-PCR assay Ct values (Ct: 14.4 - 27.49) and one LASV positive sample that was not tested by Altona real-time qRT-PCR assay. To measure the proportion of LASV nucleic acid present, assess the genome coverage and to optimise the MinION data analysis pipeline by comparing the consensus sequences generated with both the MinION (Oxford Nanopore) and MiSeq (Illumina) platform, all samples were processed as described in Section 2.8, Nanopore sequenced as described in Section 2.10 and MiSeq sequenced as described in Section 2.9 (Table 5.4). Reads were *de novo* assembled following data processing as described in Section 2.11 and Section 2.20.3 and contigs generated were used for the identification of appropriate LASV reference sequences for each sample.

Table 5.4 Description of samples positive for LASV by Altona real-time reverse transcription-PCR and by Nikisins real-time reverse transcription-PCR with corresponding MinION and MiSeq reads, along with total percentage of reads mapping to LASV for each platform (n = 14 samples).

Sample ID	Altona Ct Value	Nikisins Ct Value	MinION Total Reads	MiSeq Total Reads (R1+R2)[a]	MinION % LASV	MiSeq % LASV
ISTH-2018-073	14.4	21.18	368369	4048096	26.74	26.17
ISTH-2018-119	17.33	16.36	1284263	3672114	24.32	29.49
ISTH-2018-126	17.96	20.5	1121865	3615248	16.3	20.44
ISTH-2018-066	18.74	26.73	84401	3992656	36.46	32.63
ISTH-2018-014	19.16	26.42	887781	3540100	4.66	10.13
ISTH-2018-131	20.57	25.68	829314	3557916	2.15	1.68
ISTH-2018-115	21.00	26.55	509947	3436612	3.09	3.66
ISTH-2018-075	22.39	29.74	554028	7015848*	2.45	0.12
ISTH-2018-072	22.63	34.3	353084	6756916*	16.48	0.22
ISTH-2018-021	23.82	31.76	571510	8340252	2.97	2.94
ISTH-2018-013	24.01	26.02	2478021	7384854*	3.14	1.2
ISTH-2018-074	26.87	32.9	630833	6581500*	0.69	0.1
ISTH-2018-001	27.49	23.75	1468422	3962172	7.23	3.24
ISTH-2018-036	None	28.44	218135	7186332*	2.38	0.23

[b] 'R1+R2' indicates paired-end sequencing

*Samples were run on two separate MiSeq sequencing runs and reads were combined prior to data analysis

For both segments combined the total percentage of reads mapping to LASV were comparable with the majority of samples presenting with less than 5% difference between the two platforms (Table 5.4). Sample ISTH-2018-72 (Altona Ct: 22.63) was the only sample to present with a percentage difference of reads mapping to LASV above 10%, with 16.48% for the MinION and 0.22% for MiSeq. The least difference was observed for sample ISTH-18-021 (Altona Ct: 23.82), with 2.97% for MinION and 2.94% for MiSeq. Figure 5.2 shows the percentage of reads mapping to the respective viral reference for both platforms used and for each segment separately, additional information with the 20x percentage genome coverage succeeded for each can be found in Figure S2, Table S1 and Table S2. Genome recovery of >95% was succeeded for all MinION generated genomes and for all MiSeq ones except the L segment sequences of sample ISTH-2018-75, ISTH-2018-72 and ISTH-2018-074 which had a 20x genome coverage of 87.43%, 87.2% and 72.69% respectively.

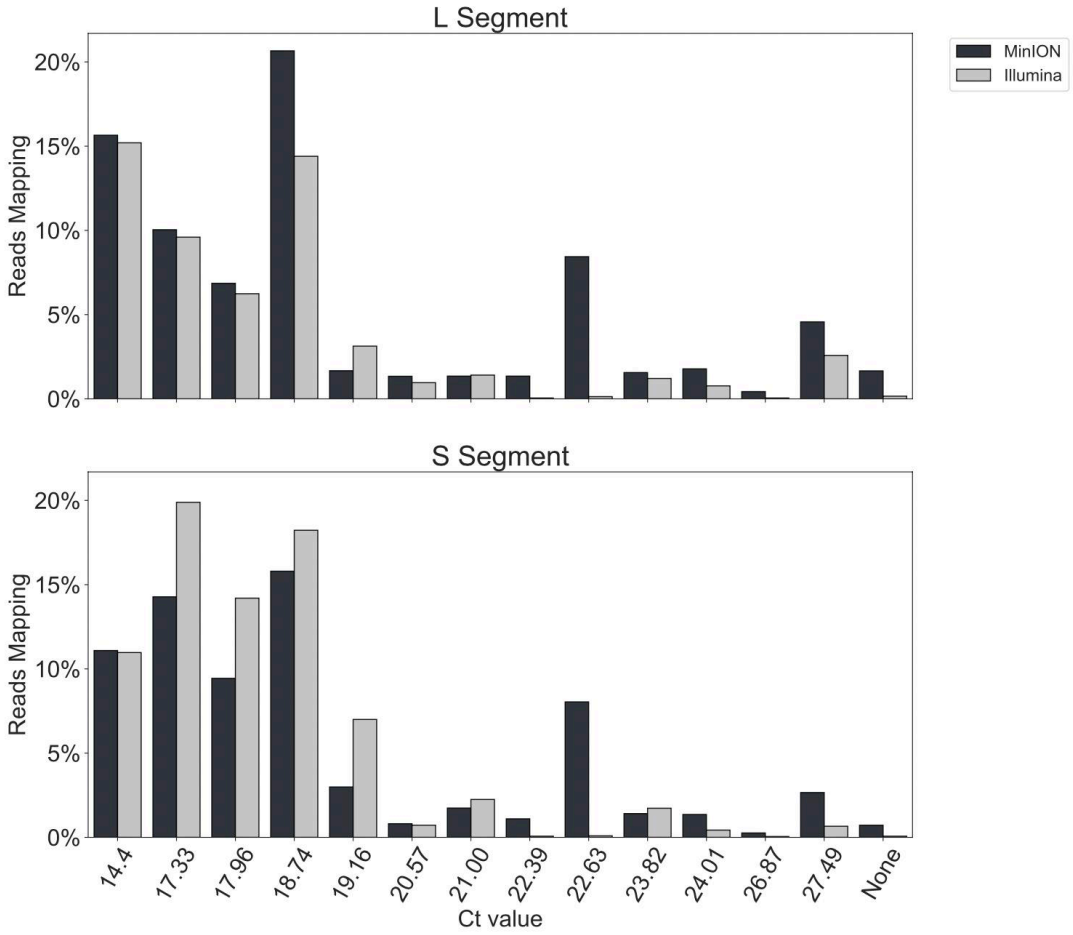


Figure 5.2 Proportion of reads mapping to the appropriate viral reference for each segment separately (n = 14 samples).

The percentage of total reads mapping to each segment is plotted for each sample and for both platforms.

Consensus sequences were generated for all MiSeq samples as described in Section 2.12.2 and for the MinION data two analysis options were investigated to identify the best approach for the generation of high-confidence consensus sequences for phylogenetic inference. Both basecalled reads and raw signal data were mapped to the reference sequence identified and variants were called using Nanopolish software, as developed for the West African Ebola virus disease outbreak (147); an additional consensus sequence was generated using the Nanopolish consensus and remapping basecalled reads to apply a further round of correction, the pileup script called bases at a minimum depth of 20x and 70% support fraction (Section 2.12.2). Consensus sequences generated from the MinION sequencing data were compared to the MiSeq generated consensus sequences. The Nanopolish MinION consensus sequences reached an accuracy level of >99% for the L Segment and >98% for the S segment and the Nanopolish/pileup consensus sequences reached ≥ 99.9%, matching their MiSeq counterparts with little to no divergence and confirming the accuracy of the Oxford Nanopore using this approach (Table 5.5, Table S3). Following these results, the MinION data analysis pipeline was finalised and Figure 5.3 presents the workflow used for the MinION consensus generation.

Table 5.5 Summary of nucleotide differences between MiSeq generated consensus and MinION generated consensus sequences using Nanopolish and Nanopolish with an additional step of correction using a voting correction

Sample ID	Altona Ct Value	Nikisins Ct Value	Segment Length	Nanopolish	Nanopolish/Pileup
L Segment					
ISTH-2018-073	14.4	21.18	7260	59	4
ISTH-2018-119	17.33	16.36	7258	49	2
ISTH-2018-126	17.96	20.5	7245	33	0
ISTH-2018-066	18.74	26.73	7135	3	1
ISTH-2018-014	19.16	26.42	7245	10	1
ISTH-2018-131	20.57	25.68	7238	0	0
ISTH-2018-115	21.00	26.55	7183	6	3
ISTH-2018-075	22.39	29.74	7256	10	0
ISTH-2018-072	22.63	34.3	7250	5	0
ISTH-2018-021	23.82	31.76	7237	5	0
ISTH-2018-013	24.01	26.02	7230	36	0
ISTH-2018-074	26.87	32.9	7238	9	0
ISTH-2018-001	27.49	23.75	7196	37	2
ISTH-2018-036	None	28.44	7261	7	1
S Segment					
ISTH-2018-073	14.4	21.18	3406	36	1
ISTH-2018-119	17.33	16.36	3403	13	2
ISTH-2018-126	17.96	20.5	3407	2	1
ISTH-2018-066	18.74	26.73	3393	21	0
ISTH-2018-014	19.16	26.42	3407	0	1
ISTH-2018-131	20.57	25.68	3389	5	0
ISTH-2018-115	21.00	26.55	3387	0	0
ISTH-2018-075	22.39	29.74	3412	1	0
ISTH-2018-072	22.63	34.3	3398	0	6
ISTH-2018-021	23.82	31.76	3385	4	0
ISTH-2018-013	24.01	26.02	3367	9	0
ISTH-2018-074	26.87	32.9	3367	1	0
ISTH-2018-001	27.49	23.75	3490	0	0
ISTH-2018-036	None	28.44	3387	2	0

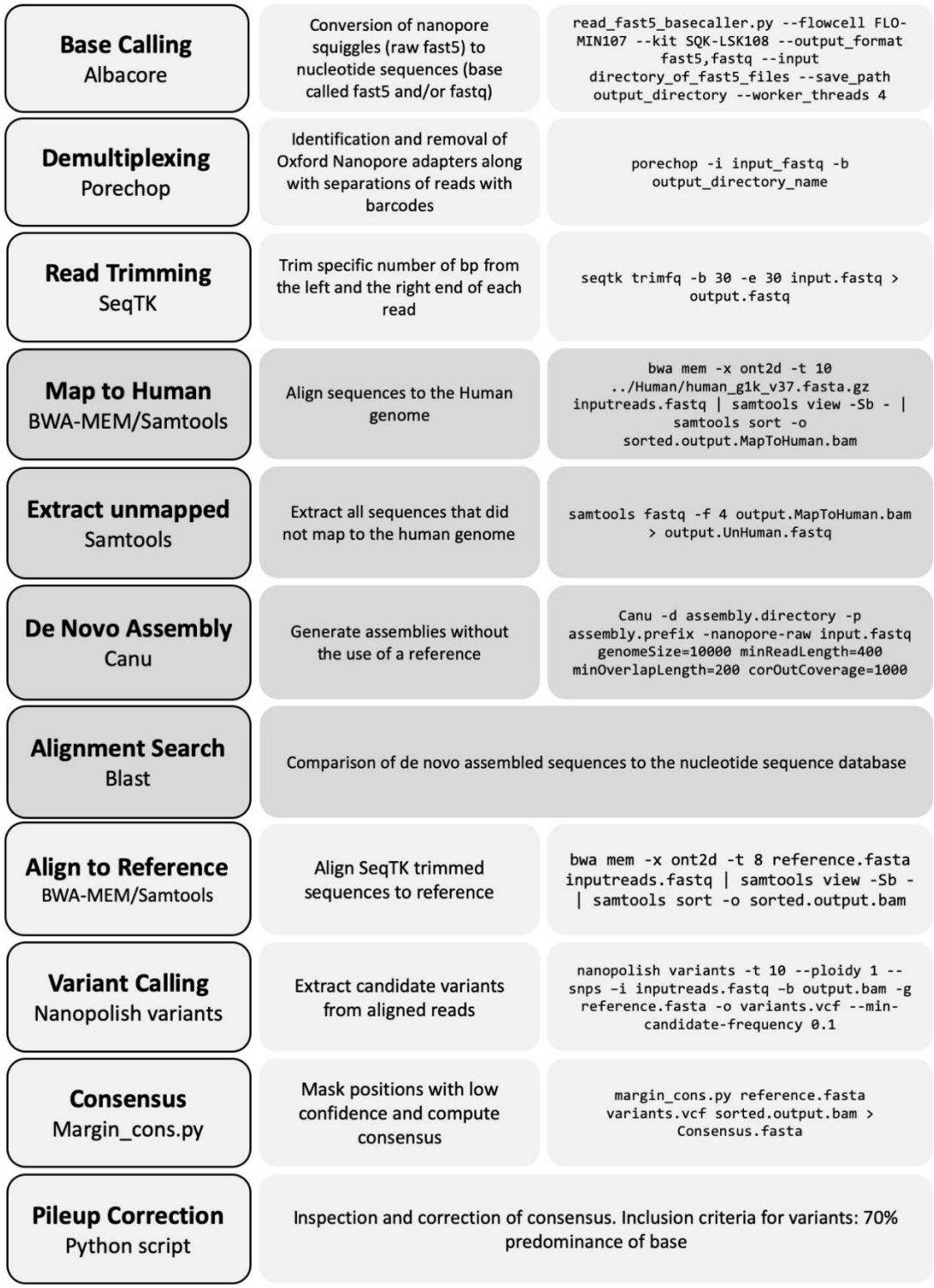


Figure 5.3 Workflow of consensus sequence generation.

Summary of the steps performed during the bioinformatics pipeline for consensus generation.

5.3.3 Metagenomic MinION Sequencing in a resource-limited setting

A total of 120 samples were sequenced, selected based on Ct value and location from the 341 positive cases reported by ISTH between 1st January and 18th March 2018. Samples tested positive for LASV by either or both Altona real-time RT-PCR (targeting S segment) and Nikisins real-time RT-PCR (targeting L segment) and represented a diverse range of Ct values. Altona Ct values ranged from 14.4 to 37.87 and Nikisins Ct values from 16.36 to 41.32 (Figure 5.4). Selected samples covered the wide range of clinical viral loads observed, including several samples testing negative in one of the two real-time RT-PCR assays used.

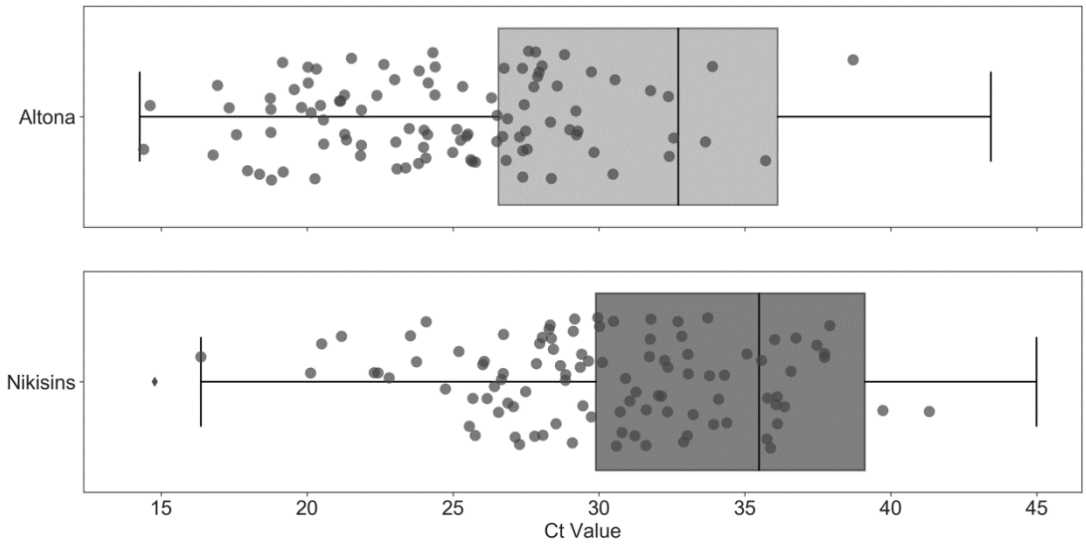


Figure 5.4 Cycle threshold values distribution of LASV samples tested positive by the Altona assay (n = 286) and the Nikisins assay (n = 228) from the Institute of Lassa Fever, Research and Control, Irrua Specialist Teaching Hospital (n = 341 first test samples).
First test samples sequenced (total = 105) are indicated by circles, 97 samples were positive by the Altona assay and 101 samples were positive by the Nikisins assay. For each assay, the median Ct value of positive samples by quantitative real-time PCR is shown (horizontal line inside box), as well as, 25th and 75th percentiles (box lower and upper boundaries) and total range (whiskers).

The virus diversity and the challenge of LASV diagnosis is highlighted when correlating the results of the two diagnostic assays used. Figure 5.5 shows the relation of Ct value distributions of the samples sequenced for both assays (n = 103 for samples that were tested by both assays and found positive in both). An average Ct value difference of 5.9 is observed between Altona Ct value and Nikisins Ct value for samples that are positive for both assays. A total of 105 samples sequenced were first test samples and the remaining 15 samples sequenced were follow-up samples. From the 120 samples sequenced, 103 were positive for both assays, 9 were not tested and one was found negative by the Altona assay and 8 were negative in the Nikisins assay (Table 5.6).

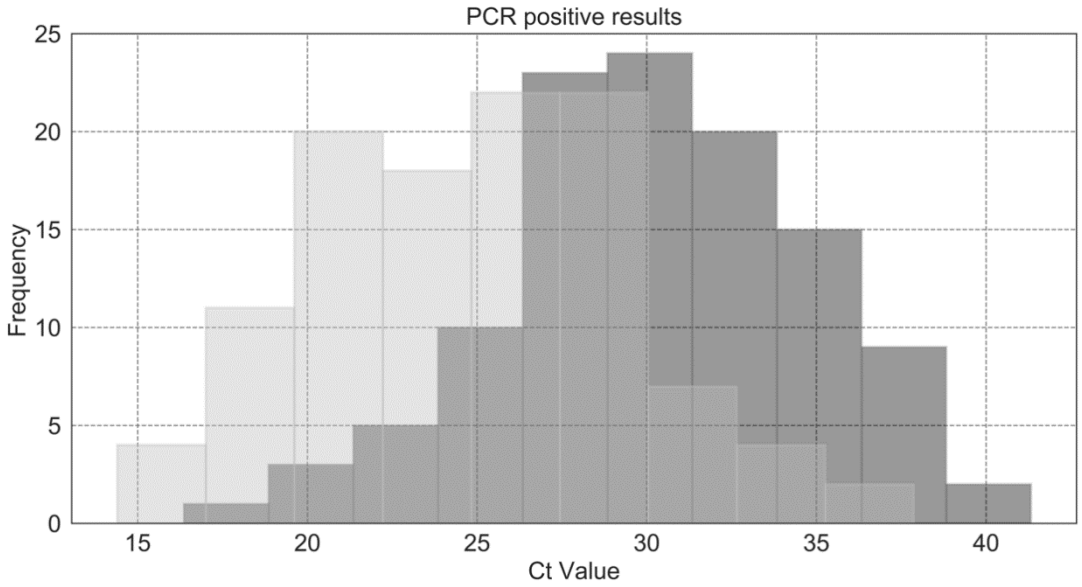


Figure 5.5 Cycle threshold value distribution of sequenced LASV positive samples from the Institute of Lassa Fever, Research and Control, Irrua Specialist Teaching Hospital (n = 120 sequenced samples).

The distribution of sequenced positive LASV samples Ct values tested with the Altona assay are depicted in light grey and with the Nikisins assay in dark grey.

Table 5.6 Description of LASV positive samples sequenced

	PCR Positive – First test		First Test Sequenced		Total Sequenced	
Sample Number	341		105		120	
Number of samples per results						
Result	Altona	Nikisins	Altona	Nikisins	Altona	Nikisins
Positive	286	228	97	101	110	112
Negative	9	111	1	4	1	8
No real-time RT-PCR results (gel-based or not done)	45	2	7	0	9	0
Number of samples per range of Ct value						
Ct value range	Altona	Nikisins	Altona	Nikisins	Altona	Nikisins
14 to 30	110	58	88	49	97	55
30 to 35	87	50	8	34	11	35
35 to 45	89	120	1	18	2	22

From the total of first sample tests concluded positive (1st of January and the 18th of March 2018), 80% (88/110) of Altona positive samples with Ct range 14 to 30 and 77% (83/108) of Nikisins positive samples with Ct range 14 to 35 were sequenced (Table 5.6). The correlation between Ct values of the positive samples sequenced for the two PCR assays is shown in Figures 5.6, one sample tested negative in the Altona assay and eight samples tested negative in the Nikisins assay, demonstrating the importance of combined use of both assays for diagnosis of acute Lassa fever and subsequent evaluation for sequencing.

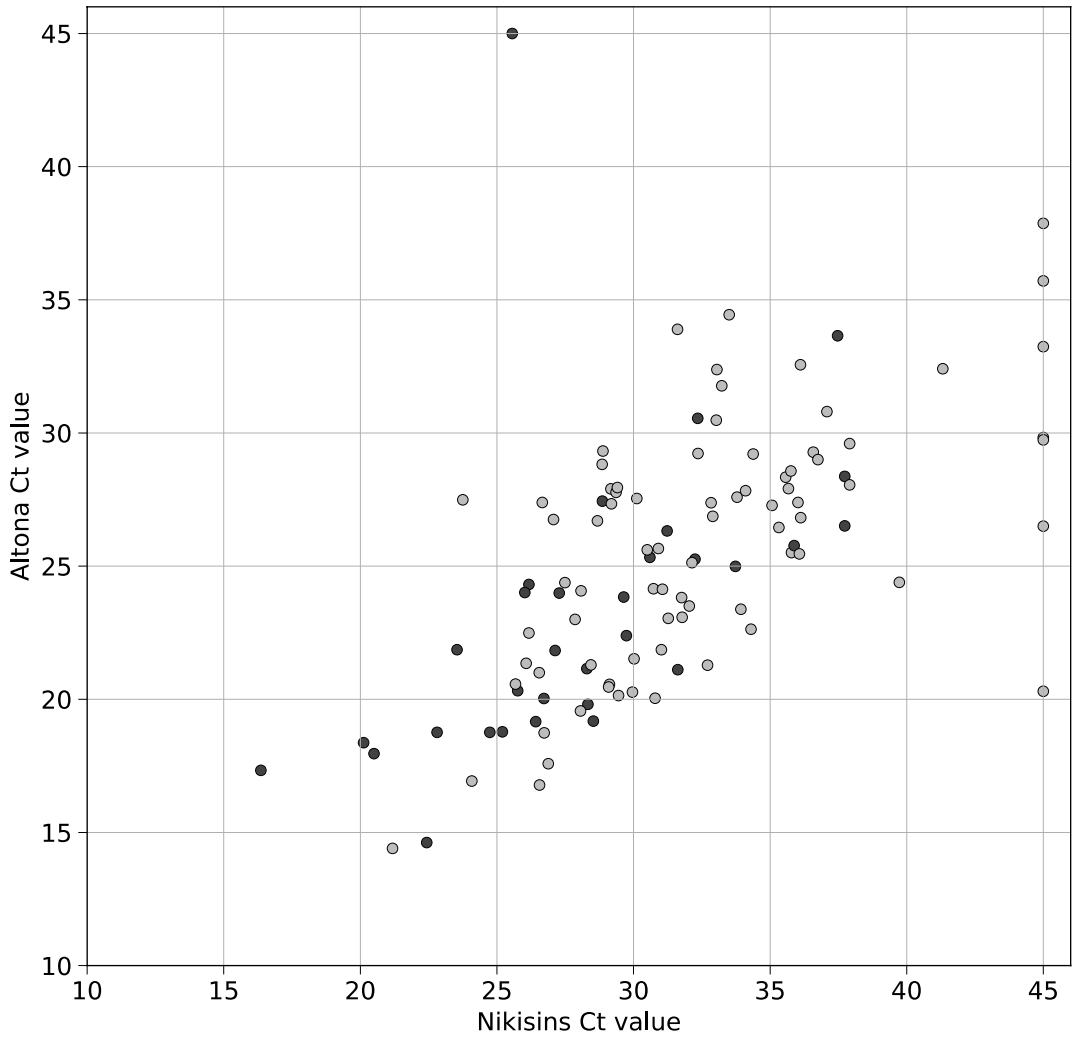


Figure 5.6 Correlation between Ct values from Altona and Nikisins real-time RT-PCR assays.

Scatter plot correlating the two assays and including categorisation of the data based on patient outcome. Values for samples from survivors are plotted in grey, and those from deceased patients in black. One sample tested negative in the Altona assay and eight samples tested negative in the Nikisins assay. Negative results have been assigned a Ct value of 45 to facilitate visualisation.

Samples selected showed an even distribution over the different states affected during the endemic season. Total first samples tested at ISTH during the period of 1st January until the 18th of March and their distribution across the different states can be found in Table 5.7. The samples we sequenced correlate with this distribution (Table 5.7). Figure 5.7 shows the sample distribution over the different states with a choropleth representation of the total positive LASV cases during the period we were in-country and orange markers depicting the location of the samples sequenced. The majority of the samples originated from Edo state followed by Ondo and Ebunyi (Table 5.7, Figure 5.7)

Table 5.7 Distribution of first test positive samples and samples sequenced across the different states.

State	First test positive samples (n = 341)	Samples Sequenced (n = 105)
Taraba	1	0
Benue	2	0
Ondo	95	29
Delta	7	2
Abia	1	0
Ebonyi	42	14
Anambra	6	5
Edo	162	42
Bauchi	6	5
Nasarawa	1	1
Kogi	10	4
Imo	4	2
Ekiti	2	1
Federal Capital Territory	2	0

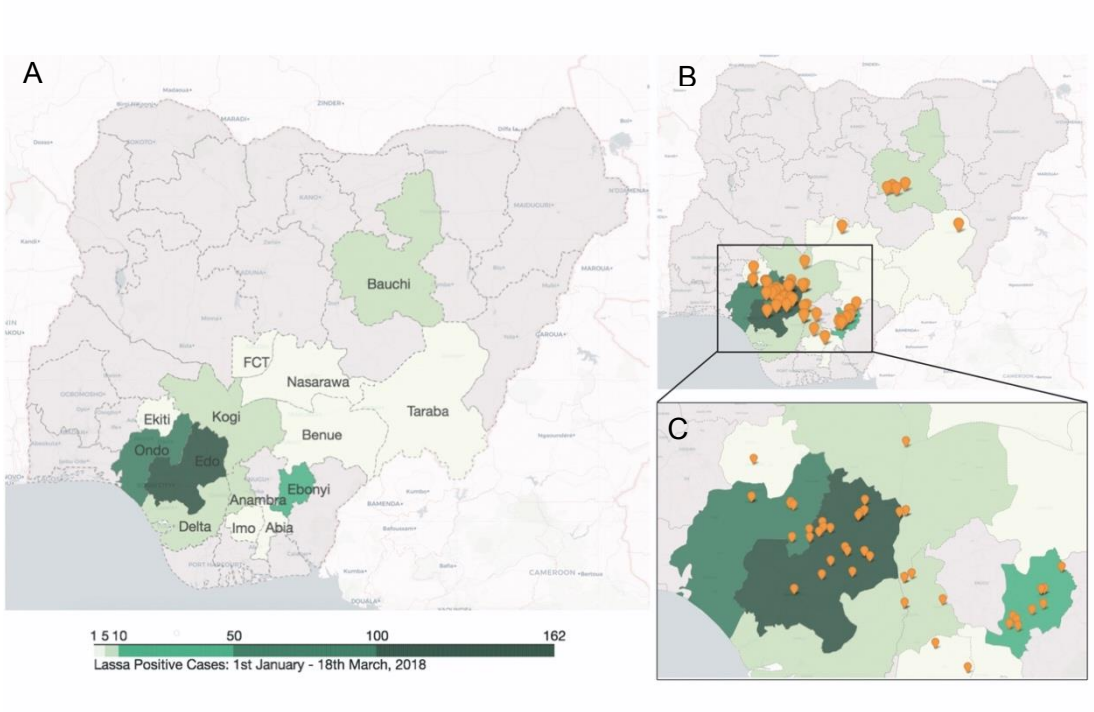


Figure 5.7 Annotated map of confirmed Lassa fever cases between 1st January and 18th of March and of samples sequenced

(A) Affected states and choropleth annotation per number of cases (B-C) geographical origin of patients from whom samples were sequenced (orange markers)

A notable proportion of total reads generated per sample were LASV (L and S segment) at an average frequency of 4.26% with a maximum of 42.9% (Figure 5.8). Average frequency for the L segment was 2.09% and for the S segment 2,16% and a maximum of 20.60% and 24.29% was observed for each segment respectively (Figure 5.9), allowing for sufficient genomic sequence (>70%) for phylogenetic comparison of at least one segment in 91 of the samples tested (Figure 5.10).

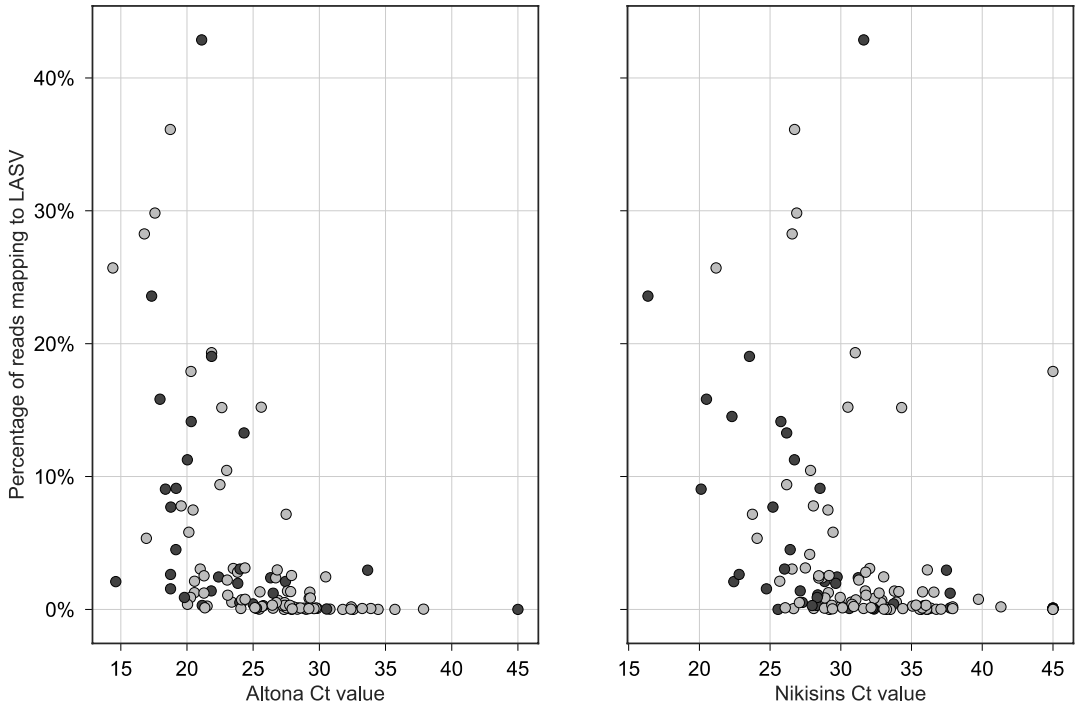


Figure 5.8 Percentage of reads mapping to LASV (L and S segment) depending on Ct value in Altona and Nikisins real-time RT-PCR assay.

Negative results have been assigned a Ct value of 45 to facilitate visualisation. Values of samples from survivors are plotted in grey and those from deceased patients are in black.

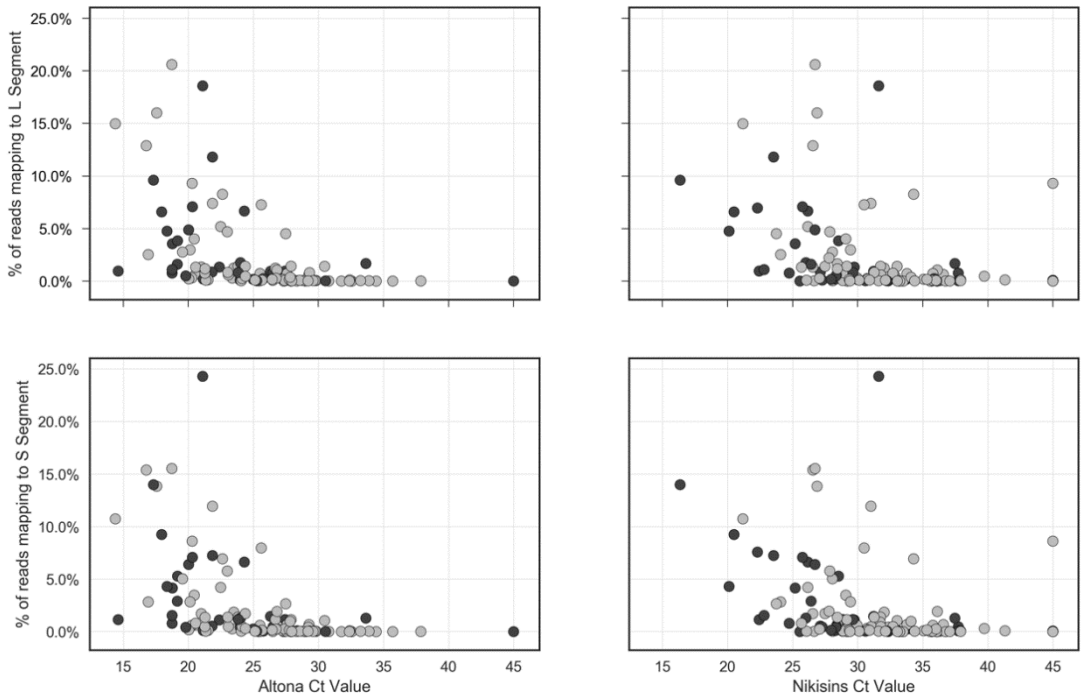


Figure 5.9 Percentage of reads mapping to each LASV segment depending on Ct value in Altona and Nikisins real-time RT-PCR assay.

The percentage of total reads mapping to the L segment id plotted in the upper panel and to the S segment in the lower panel. Values of samples from survivors are plotted in grey and those from deceased patients are in black.

The proportion of total reads mapping to LASV reference with a minimum depth of 20 reads (20x) is shown in Figure 5.7 for both segments and both assays. The pattern is more consistent with samples presenting with a Ct value of 25 and below in Altona and Ct value of 30 and below of Nikisins are more consistently expected to return sufficient reads for 20x coverage of the reference genome.

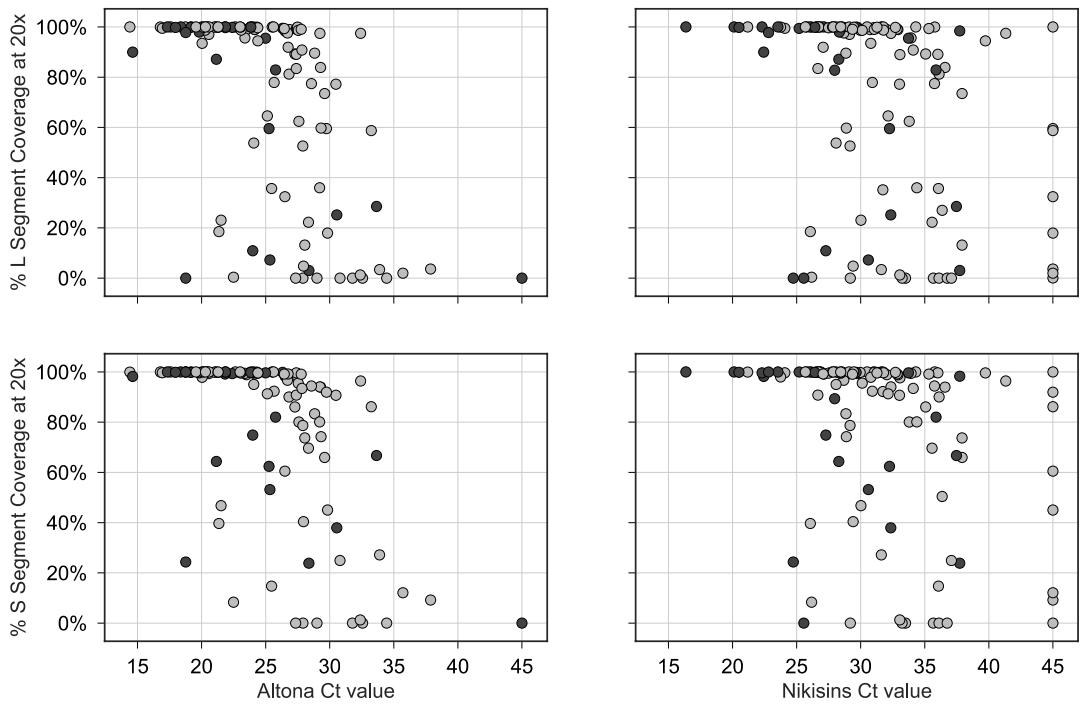


Figure 5.10 The proportion of total reads mapping to LASV L segment and S segment with a minimum depth of 20 reads (20x) depending on Ct value in Altona and Nikisins real-time RT-PCR assay.

The percentage of total reads mapping to the L segment is plotted in the upper panel and to the S segment in the lower panel. Values of samples from survivors are plotted in grey and those from deceased patients are in black.

Metagenomic classification using the Centrifuge software (*326*) identified 0.10% of reads from sample 110 as originating from hepatitis A virus, with read mapping providing 74% genome coverage at 20-fold depth. LASV accounted for 0.83% of reads in the same sample, providing 96% genome coverage. These findings demonstrate the potential of this simple approach to identify multiple RNA viruses, including those present as co-infections. In all other samples tested, LASV was the sole pathogen identified despite a small number of reads classified as other viruses (Figure 5.11a and Figure 5.11b).

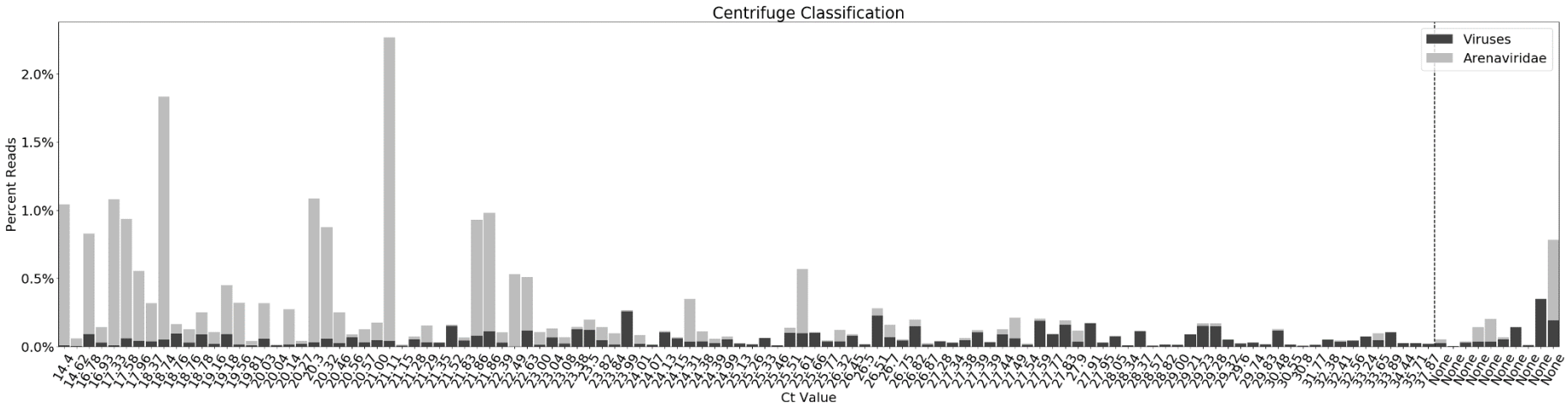


Figure 5.11a Classification of MinION reads depending on Ct value in Altona.

Reads were classified by Centrifuge software as either Arenaviridae or other viruses. The analysis allowed for identification of a single coinfection in sample 110 with 0.1% reads classifying as Hepatitis A virus. In all other samples, the distribution of reads classified 5 within the other viruses did not include a sufficient proportion of specific origin to suggest the presence of a virus other than LASV.

None: no Ct value as samples were not tested with the respective RT-PCR

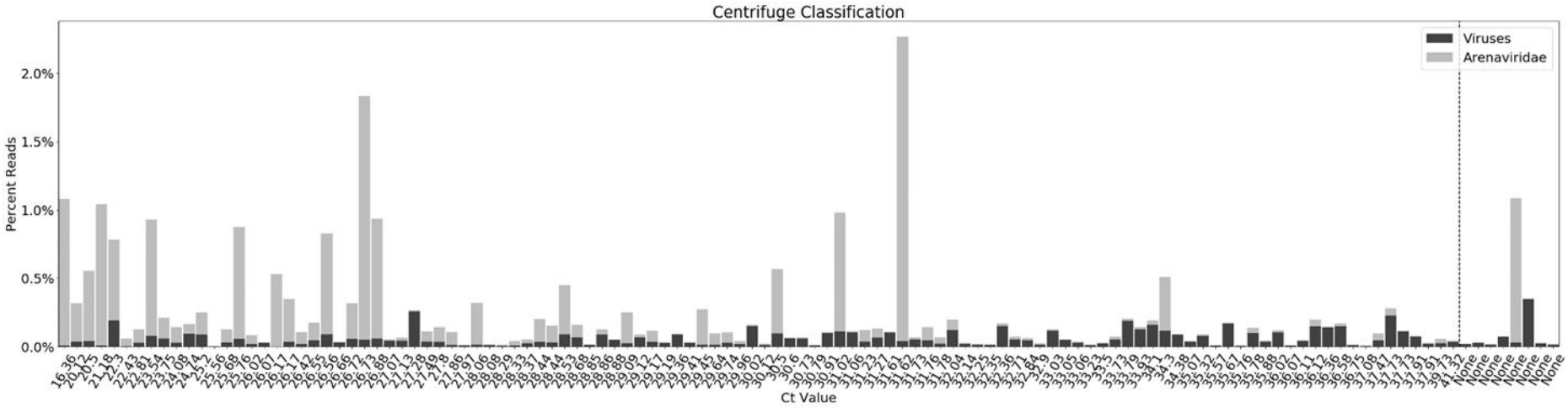


Figure 5.11b Classification of MinION reads depending on Ct value in Nikisins real-time RT-PCR assay.

Reads were classified by Centrifuge software as either *Arenaviridae* or other viruses. The analysis allowed for identification of a single coinfection in sample 110 with 0.1% reads classifying as Hepatitis A virus. In all other samples, the distribution of reads classified 5 within the other viruses did not include a sufficient proportion of specific origin to suggest the presence of a virus other than LASV.

None: no Ct value as samples were not tested with the respective RT-PCR

5.3.4 Metagenomic MinION sequencing during the 2018 outbreak

The upsurge of Lassa fever cases during the 2018 endemic season in Nigeria was the largest on record, reaching 1,495 suspected cases and 376 confirmed cases and affecting more than 18 states by 18 March (Figure 5.12). This notably exceeds the 102 confirmed cases reported during the same period in 2017 (Figure 5.12) (*340*). The unprecedented scale of the outbreak raised fears of the emergence of a strain with a higher rate of transmission. Due to these concerns, on 28 February the Nigeria Centre for Disease Control (NCDC) and the WHO urgently requested sequencing information and preliminary results from our pilot-scale study, in which the metagenomic approach with the MinION device (Oxford Nanopore Technologies) was used to conduct in-country, mid-outbreak viral genome sequencing. This instigated a major increase in sequencing efforts, leading to the sequencing of 120 samples.

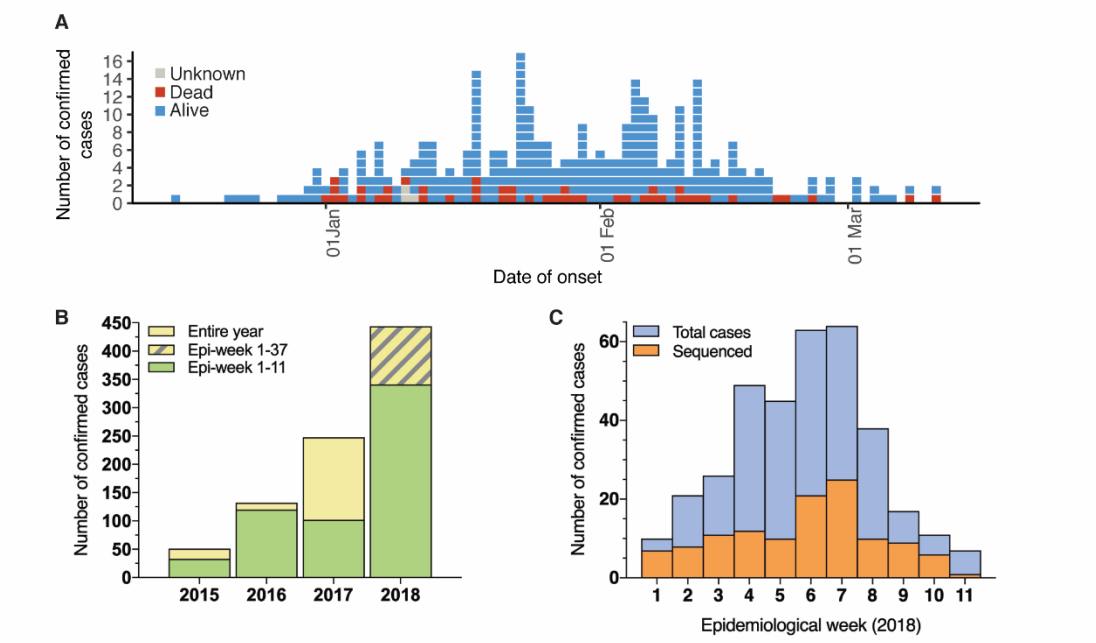


Figure 5.12 Epidemiology of the Lassa fever outbreak and timeline of sequencing in Nigeria.

Figure taken from Kafetzopoulou et.al. (303) and generated by S. Duraffour

(A) Epidemiological curve for 2018. ISTH confirmed 341 of the 376 Lassa fever cases reported by Nigeria Center of Disease Control (NCDC) between 1st January and 18th March 2018. The epidemiological curve shows the 341 confirmed cases according to patient outcome. (B) Number of cases diagnosed and reported by ISTH from 2015 through 2018. (C) Number of samples sequenced per epidemiological week in 2018.

The request from NCDC for information on circulating strains was made on 28 February at the height of the outbreak; within 10 days, our pilot study was expedited and the initial analysis completed. The total 120 LASV-positive samples selected on the basis of cycle threshold value and location of the 341 cases reported by ISTH between 1 January and 18 March 2018 were sequenced during a 7-week mission. Rapid sequencing and real-time analysis of 35 genomes revealed sequence diversity, suggesting independent zoonotic transmission events as the cause; allaying concerns of an emergent novel strain. Subsequent analysis of 85 further samples sequenced during the outbreak confirmed these findings highlighting the power of real-time metagenomic sequencing in outbreaks. The fact that the 2018 outbreak was fueled by the circulating LASV diversity and not by transmission of a new or divergent lineage was already evident from the first seven genomes generated by 10 March (Figure 5.13).

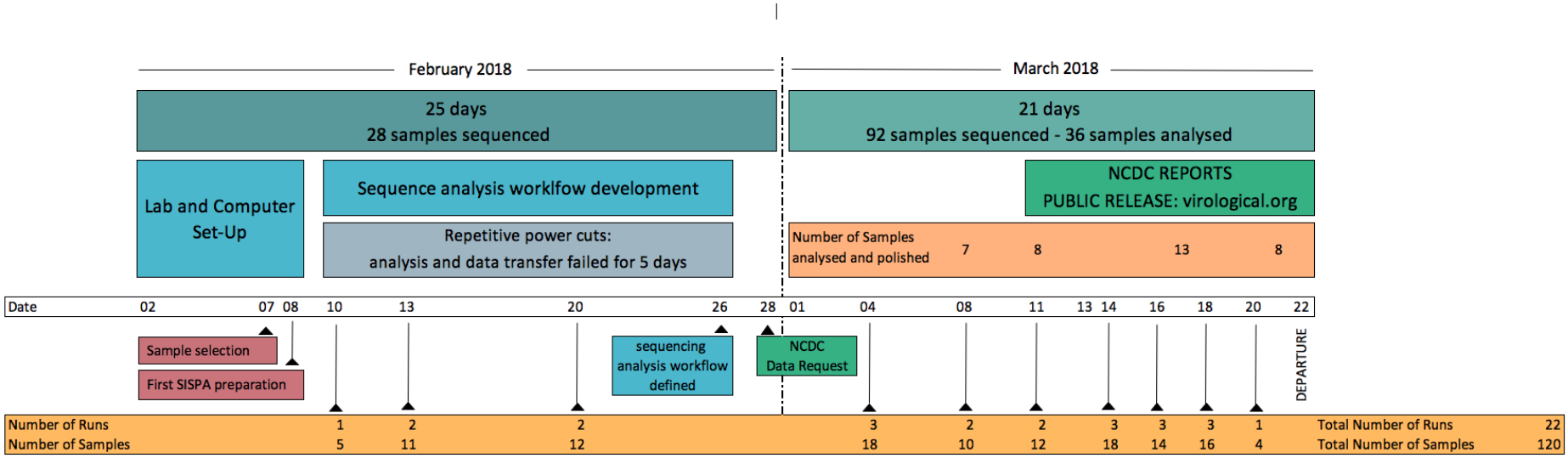


Figure 5.13 Timeline of sequencing in Nigeria.

Timeline of sequencing efforts. Equipment and consumables for sequencing of ~50 samples and the computer hardware were deployed at ISTH with the aim of testing and troubleshooting on-site sequencing capacity. Sequencing data was requested by the NCDC on the 28th of February. The alarming increase in cases effectuated an upscale in efforts leading to sequencing of 120 samples on-site.

This information was promptly communicated to the NCDC, forming the basis of its report released on 12 March 2018 (*341*). Whereas this first small sample was restricted to genotype II, the final collection of 36 LASV genome sequences generated on-site also included a representative of genotype III (Figure 5.14), further supporting the spillover of long-standing LASV diversity in the outbreak. Consensus sequences generated on-site were communicated in four different releases, with the first sequences from 7 samples released online on March 12th (*342*), the following from 8 samples on March 15th (*343*), an additional 12 on the 21st of March and the final report on the 4th of April with the total 35 samples sequenced and analysed on site (*343*).

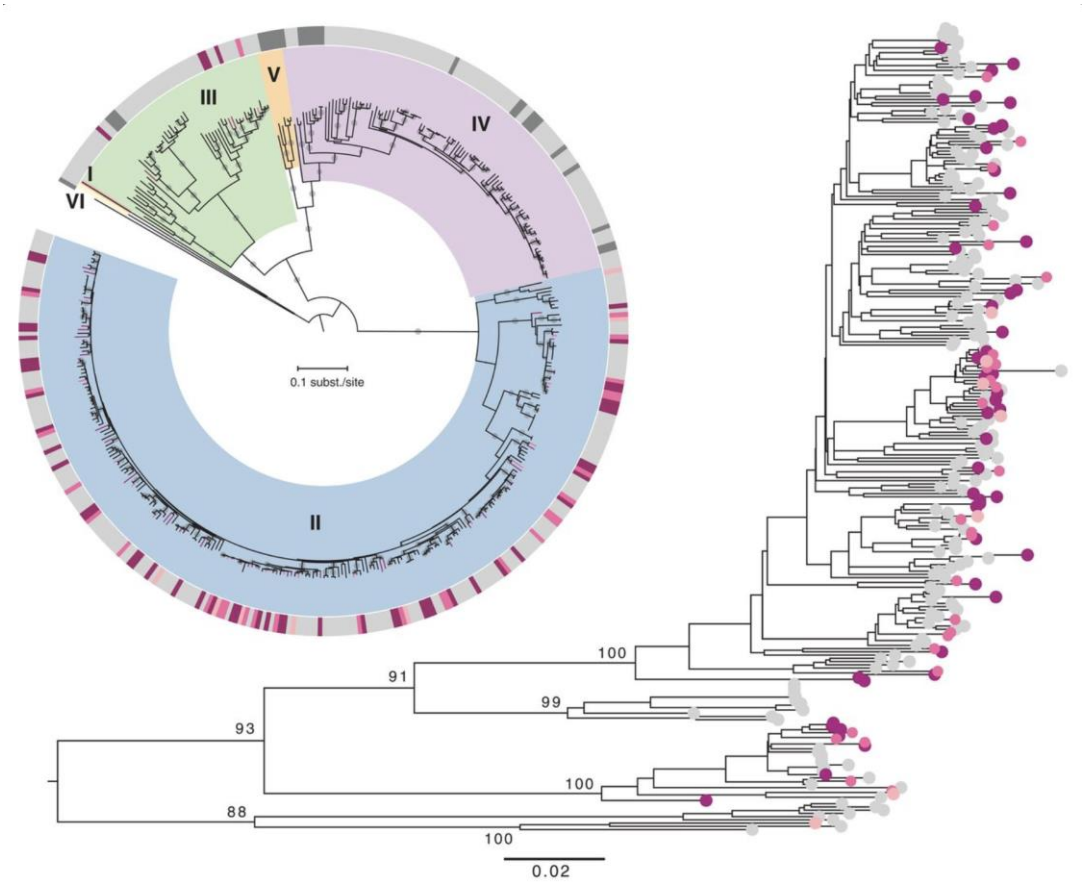


Figure 5.14 Phylogenetic reconstruction of the S segment data.

(Figure taken from Kafetzopoulou et.al. (303) and generated by P. Lemey).

The circular tree includes 96 sequences from 2012 to 2017, 88 sequences from 2018, and sequences available from GenBank. The rectangular tree focuses on the genotype II clade (in blue in the circular tree), which includes most of the 2018 sequences. The six genotypes are indicated with different colours and roman numerals. Bootstrap support >90% is indicated with a small grey circle at the middle of their respective branches. The colour strip highlights the human LASV sequences obtained from previous years (light grey); sequences obtained from rodent samples (dark grey); and, for 2018, the first seven sequences generated in Nigeria (light pink), the remaining 28 sequences analysed on-site (medium pink), and the remaining sequences finalized in Europe (dark pink). The same colour code is used in the genotype II rectangular tree. Bootstrap values >80% are shown for the major genotype II lineages.

5.4 Discussion

5.4.1 Pre-deployment testing

The level of viral reads corresponding to LASV are significantly lower compared to the ones observed for CHIKV and DENV positive samples (Chapter 4), nevertheless metagenomic MiSeq sequencing of the 15 historical LASV samples (Section 5.3.1) achieved genome coverage (20x) of >80% for 7 samples for the L segment and 9 samples for the S segment. A genome recovery of >50% was successful for 9 samples for the L segment and 12 samples for the S segment. The Ct values do not directly correlate with the percentage of reads mapping to LASV for each sample and despite the low percent of viral reads observed, successful recovery of more than 50% of the genome was achieved for 80% of the samples sequenced for at least one segment and 60% for both segments (Figure 5.2). The principal limitation of the metagenomic approach is the limit of detection, however the results presented clearly show the benefit of using this approach for a genetically diverse RNA virus, particularly when the diagnostic assay sensitivity is variable due to the high inter-strain nucleic acid sequence variation of LASV. Sequencing a Ct value range of LASV positive samples allowed for the generation of genomic sequences for the majority of the samples tested and provided information on the variation in the proportion of reads corresponding to LASV for each sample. Notably samples LASV04 (Ct:18.66), LASV07 (Ct: 23.07) and LASV11 (Ct:28.97) with mid-to-high Ct values exhibit higher % of LASV than expected with 4.10%, 6.27% and 12.69% of reads respectively corresponding to LASV. The high degree of genetic variability of the virus, with strain variation up to 32% and 25% for the L and S segment respectively (*48*), creates challenges for its detection and can explain the unexpected pattern in read mapping observed. The application of metagenomic sequencing has allowed for sufficient genome recovery from the majority of the samples sequenced which can provide useful information for assay optimisation and further understanding of the virus diversity.

5.4.2 Nanopore pipeline testing and validation

Cross platform comparison of the 14 positive samples randomly selected from the 2018 Nigeria outbreak cohort allowed for the refinement of the MinION analysis pipeline. The total percentage of reads mapping to LASV was comparable between both sequencing platforms, with the majority of samples presenting a cross-platform

difference of less than 5% (Table 5.4, Figure 5.2). Genome recovery of >95% was successful for all S segments and of >70% for all L segments allowing for sufficient genomic sequence recovery (>70%) for sequence comparison (Figure S2, Table S1, Table S2). Consensus sequences generated from both platforms were compared, the Nanopolish MinION consensus reached an identity level of ≥99% for the L Segment and >98% for the S segment. The Nanopolish/pileup consensus sequences were ≥99.9% identical to their illumina counterpart with the majority of sequences reaching 100% or having only one nucleotide difference (no differences: 16/28, single nucleotide difference:6/28). Nanopolish/pileup consensus sequences matched their MiSeq counterparts with little to no divergence, confirming the accuracy of the Oxford Nanopore using this approach. The nucleotide differences identified between the MiSeq and the MinION consensus sequences were further interrogated to identify the nucleotide disagreements within the coding regions, which were used in phylogenetic analysis. Ten of 14 had zero differences in both S or L segment coding regions, whilst four had 1-3 nucleotide disagreements in total across the combined S and L coding regions. Visual inspection of these regions suggested the basecalling was consistent within the read alignment for both the Illumina and Nanopore data and so does not appear to be the result of the extra “noise” within the Nanopore signal (Table S4). Integrating the additional correction step post Nanopolish generated the most accurate consensus, leading to the refined MinION analysis pipeline used (Figure 5.3).

5.4.3 In-country sequencing

Demonstrating the utility of metagenomic sequencing for LASV patient samples, allowed for the design of a pilot study to evaluate metagenomic sequencing in a resource-limited setting using the Oxford Nanopore MinION. Metagenomic MinION results clearly show that the amount of viral reads generated are sufficient to successfully generate near complete viral genomes for both segments. More than 70% of each segment was recovered for the majority of the 120 samples sequenced, demonstrating that the metagenomic sequencing approach employed is capable of elucidating significant portions of viral genome of LASV, a highly divergent RNA virus. Ct value does not directly correlate with the proportion of sequencing reads observed, presenting with a considerable level of variation between samples, especially within the Ct range of 20 to 30 in the Altona assay and 25 to 35 in the Nikisins assay (Figure 5.9). This variation is most likely attributed to the diversity of LASV and the difficulty of designing a RT-PCR which maintains a high sensitivity in detection across all the

different LASV circulating LASV. This is highlighted with sample ISTH-2018-025 which has an Altona Ct value of 21.11 and a Nikisins Ct value of 31.62 but resulted in the highest percentage of reads mapping (42.86%) to LASV observed in the whole cohort of samples sequenced. From this percentage, 18.57% of the reads were attributed to the L segment and 24.29% to the S segment, percentages observed in the majority of samples in the lower Ct value range (<20 in Altona and <25 in Nikisins) of PCR positive LASV samples. The lowest amount of viral reads observed was 0.06% for three different samples (ISTH-2018-058, ISTH-2018-080, ISTH-2018-142), yet still sufficient to generate 73.49%, 35.97% and 83.45% of the L segment and 65.95%, 80.11% and 90.78% of the S segment at 20x for each sample respectively. Only 29 samples failed to produce sufficient genomic information, the Ct value distribution for these samples ranged from 18.76 to 37.87 for the Altona assay and from 24.74 to 37.73 for the Nikisins assay, emphasizing the difficulty of using Ct value as a successful predictor of read percentage corresponding to LASV. Sufficient genomic sequence (>70%) for phylogenetic comparison was successful for at least one segment in 91 samples tested and for both segments in 76 of those samples. The Ct value range of the successfully sequenced samples ranged between 14.4 to 33.24 for the Altona assay and 16.36 to 41.32 for the Nikisins assay. Given the multiplexing of 6 samples per flow cell and an average of 990,206 reads per sample the amount of genomic information generated was highly informative and more sequencing depth of the samples with less complete consensus sequences would likely increase coverage. These results suggest that for the majority of LASV PCR positive samples (> 75%), viral load is sufficient for metagenomic sequencing directly from patient samples without further viral enrichment beyond a DNase digestion.

5.5 Conclusions

Metagenomic nanopore sequencing directly from patient samples was deployed for the first time in an outbreak epicentre, during the 2018 Lassa outbreak in Nigeria. Genomic data and phylogenetic reconstructions were communicated immediately to the Nigerian Centre for Disease Control and the World Health Organization to inform the public health response. Real-time analysis of 36 genomes and subsequent confirmation using all 120 samples sequenced in the country of origin revealed extensive diversity and phylogenetic intermingling with strains from previous years, suggesting independent zoonotic transmission events and thus allaying concerns of an emergent strain or extensive human-to-human transmission. The ruling out of case linkage and the emergence of a novel strain in real-time, during the outbreak progression was of major importance in determining the required public health response. The response was focused on intensified community engagement on rodent control, environmental sanitation, and safe food storage. Further research is needed to evaluate whether improved diagnostics and disease awareness and/or ecological and climate factors promoting transmission are the drivers behind the changing epidemiology of Lassa fever in Nigeria.

This chapter clearly highlights the importance of MinION metagenomic sequencing, which was shown here for the first time to be a feasible approach to rapidly characterise an ongoing outbreak in a resource limited setting. Therefore, portable metagenomic sequencing of genetically diverse RNA viruses, directly from patient samples, without the need to export material outside the country of origin and with no pathogen-specific enrichment, can enable real-time characterization of potential outbreaks in resource limited settings.

Chapter 6

Discussion

6. Chapter 6: Discussion

The aim of this thesis was to investigate metagenomic sequencing for the identification and complete genome recovery of pathogenic RNA viruses directly from clinical samples, with the ambition of applying it coupled with the Oxford Nanopore MinION in resource-limited settings for real-time genomic surveillance and outbreak molecular epidemiology. This was ultimately realised and led to the first ever metagenomic sequencing to be successfully performed at the heart of an on-going outbreak in real-time, in Nigeria during the 2018 Lassa virus outbreak.

6.1 Summary

Two sequence-independent methods for viral metagenomics were initially compared as described in Chapter 3 using a mock-sample and four DENV clinical samples which were sequenced on an Illumina MiSeq to assess the proportion of viral reads generated and the consensus sequence recovery. The SISPA protocol demonstrated successful viral identification and complete genome recovery for all samples tested, which in combination with its previously successful coupling with the MinION for viral detection, made it particularly promising for rapid portable metagenomic sequencing. The SISPA protocol feasibility and sensitivity for individual clinical samples were further investigated in Chapter 4 using a selection of CHIKV and DENV positive clinical samples with a representative range of viral titres (*339*). Whole genome sequencing was successful for both viruses, directly from nucleic acid extracts of serum and plasma samples without the need for culture or viral enrichment beyond a simple DNA digestion. In Chapter 5, metagenomic nanopore sequencing was evaluated in a remote and resource-limited setting. Metagenomic nanopore sequencing of LASV virus was implemented in Nigeria during the 2018 endemic season, which enabled outbreak support for LASV during the 2018 outbreak (*303*). Genomic data and phylogenetic reconstructions were communicated immediately to the NCDC and the WHO to inform the public health response (*341*).

Metagenomic sequencing was therefore demonstrated to be a feasible approach for the recovery of viral genomes directly from clinical samples for CHIKV, DENV and LASV within the clinical range of viral titres, and the feasibility of applying it in the field for real-time outbreak analysis proven.

6.2 Sequencing platforms

Advances in the field of whole genome sequencing over the last decade have allowed for genomics-informed surveillance and outbreak response. The main disadvantages of sequencing platforms for outbreak support was until recently was the size of the equipment required, the speed at which data was generated and the operation complexity, particularly when an outbreak occurs in a resource-limited setting.

The development of the MinION platform, has already introduced sequencing in real time in the field and overcomes many limitations of its predecessors (*147*, *344*), providing a highly accessible sequencing device, with low start-up costs, that can be implemented in remote locations and requires little device maintenance. One of the most important benefits of MinION sequencing, coupled to its portability, is the capacity to sequence samples locally, directly in the country of origin without the need of shipping and transporting samples to a secondary location. This is a major advantage for sample preservation (eliminating additional freeze-thaw cycles or possible damage), removes political and logistical concerns over sample transportation and more importantly allows for the establishment and advancement of research locally in resource limited locations. However despite its major portability advantage, the benefit of long reads and direct "reading" of the input DNA, it also presents quite a few challenges which are currently being addressed. These challenges include the per read accuracy of the data generated, the cost per sample sequenced. Data analysis complexity, data volume generated and data management were also difficulties that were encountered for the duration of this thesis. In Chapter 4, for example, a low number (n = 4) of samples were screened using the MinION, this was due to the lack of a compatible barcoding kit and low data output per flow cell at the time which did not allow for sample multiplexing. The cost per sample was initially much higher compared to sequencing experiments conducted later, in Chapter 5, at which stage the barcoding kit was available and the data output was significantly higher, allowing for 5 samples to be multiplexed on a single flow cell. The comparison of the Illumina generated data versus the Nanopore counterparts showed excellent agreement for the recovered consensus sequences matching with >99.9% identity, which was sufficient for the needs of this project. However this is on the consensus level and requires a minimum of 30-fold depth MinION reads to achieve (*345*), which introduces its own limitations, particularly when encountering samples that have a low viral load and subsequently generate a much lower percentage of viral reads which in multiple cases are not sufficient to call a consensus. Once

improvements are made in the library preparation and the base calling algorithms the accuracy per read and total data yield should increase, which might then allow for an improved recovery of near-complete (>70%) consensus genomes from the total samples sequenced.

The main difficulties encountered during this work were due to challenges presented when using a new and fast growing technology with ongoing improvement and upgrades. Software updates, need for internet connection and consumable changes were all aspects that had to be considered for the duration of this project. An important challenge was data analysis and management. The data load produced initially, at the time of Chapter 3, was considerable but still easily manageable. However, by the time of the experiments in Chapter 5, the data output was at least 5x more and the experiments were generating hundreds of gigabytes worth of data. This was of course a great improvement in kit and flow cell chemistry but basecalling was still at this stage done using CPU power which meant an average of 1 week to basecall a full sequencing run and a big data storage requirement to make sure the data is saved and backed up appropriately. On the whole, these challenges were manageable but added a level of complexity, particularly compared to other commonly used sequencing platforms which have been in the market for longer and subsequently are more established.

6.3 Sequencing approaches

Amplicon sequencing has been used in combination with Nanopore and Illumina platforms to successfully sequencing EBOV (*147*), ZIKV (*156*, *218*), YVF (*219*) and WNV (*219*, *220*) for virus surveillance and outbreak support. Amplicon sequencing provides a sensitive cost effective approach and is particularly important for samples with low viral titre. It has subsequently been used to successfully sequence single lineage outbreaks, where the viral pathogen is previously known and conserved. However, despite its cost and sensitivity amplicon sequencing presents a lot of limitations, particularly in the cases of co-circulating pathogens, co-infections and highly divergent pathogens. The utility of a metagenomic approach is already highlighted in Chapter 4, in which the metagenomic protocol used was successfully applied for samples with different DENV serotypes and more importantly was able to identify a co-infection. Both of these would have not been possible via targeted methods without further screening or without using multiple primer pools to assure coverage of all four DENV serotypes. The utility of metagenomic sequencing was

further highlighted in this work through its application for sequencing a divergent pathogen as described in Chapter 5. This would not have been possible without the use of a metagenomic approach, as multiple primers sets would have to be designed for each clade subset and screening would have been much more labour intensive and expensive. Metagenomic sequencing of course has its own difficulties and limitations particularly due to a more complex data analysis requirement. This was the case for LASV, as the data generated has to be analysed using an approach incorporating an initial *de novo* assembly to allow for the identification of the appropriate reference to be used downstream for the consensus generation. The additional complexity of the analysis adds mainly to the sample-receipt-to-consensus sequence time, which in this case given the utility of the information generated is a small price to pay.

6.4 Clinical Metagenomics

Clinical metagenomics, which specifically refers to the application of metagenomics directly to clinical samples, is identified here as a suitable approach for the whole genome sequencing of three pathogenic RNA viruses of major importance (CHIKV, DENV and LASV). However, despite the successful application for these pathogens and the significant benefit of non-targeted sequencing, the extensive application of clinical metagenomics is hindered by the challenges it presents. Its sensitivity depends on the sample type, background (i.e. any non target) nucleic acid abundance and the level of pathogen present which is highly variable between samples; additionally the analysis sensitivity for the data generated is also restricted by the tools and databases available (*160*). The majority of reads generated from a sample correspond in most cases to the host, which remains to date one of the biggest challenges for clinical metagenomics. This did not seem to be the case for the low Ct value CHIKV and DENV samples but was definitely the case for all LASV samples (Chapter 4 and Chapter 5). Host background along with environmental and kit contaminants (*102*) can compromise viral detection as they dominate the sample composition, this was particularly the case with LASV sample for which only one sample was identified with as high as 40% of the reads corresponding to LASV and all remaining samples had a much lower percentage of reads mapping to the virus.

6.5 In-country Sequencing

Emerging viruses are causing debilitating diseases affecting the human population and creating devastating effects globally. Their expansion in prevalence and incidence has led these viral pathogens to cause outbreaks, become endemic and spread to different continents causing increase in mortality and morbidity rates. Rapid and unbiased methods in pathogen surveillance are vital when developing a strategy for the identification, eradication and treatment of an emerging virus; more importantly in the containment of an outbreak. Genomic epidemiology in particular can provide important information to assist with knowledge gaps present when encountering emerging pathogens and their occurring outbreaks. Sequencing of emerging viruses and their circulating strains can locate their source and characterise the virus to identify any genetic changes that might have driven their emergence (*73*). Understanding chains of infections and transmission can assist in spread reduction and outbreak management. Sequencing can provide essential information particularly during an outbreak for the understanding of viral epidemiology and for the prompt identification of human-to-human transmission chains. The utility of metagenomic sequencing for genomic epidemiology and outbreak response was highlighted with the implementation of real-time metagenomic MinION sequencing during the 2018 LASV outbreak in Nigeria. The sequencing results allowed public health authorities to rule out the emergence of a new strain or the possibility of increased human-to-human transmission. Sequences and phylogenetic reconstructions were directly communicated to the Nigerian authorities and WHO allowing for the suitable allocation of resources for the informed public health response (Figure 6.1).

Figure 6.1 Sequencing report released by the Nigerian Centre for Disease Control. Early Results of LASV Sequencing & Implications for Current Outbreak Response in Nigeria (341). Full report is included in the supplementary.

A mechanism for data and information sharing was formulated consisting of an internal report directly to the NCDC and the WHO and a simultaneous public release was made available online on virological.org (*342*). Data sharing is critical during outbreaks and epidemics allowing for input from the scientific community to provide additional expertise and the possibility for global pathogen surveillance. MinION genome sequencing utilising a metagenomic approach has the potential for both rapid viral identification and genomic level data acquisition. Shown in this work for the first time in the context of an unfolding LASV outbreak, during the 2018 Nigerian endemic season, is the application of metagenomic MinION sequencing for the generation of real-time data used for downstream analysis in virus surveillance, epidemiology and viral evolution. Additionally, the sequencing information generated provides an important impact on design and improvement of diagnostics, particularly so given the LASV diversity. Currently positive diagnosis of the LASV is done using two PCR assays, due to its diversity positive cases can be missed with a single assay and using two assays increases the sensitivity of detection. The sequences generated can be used to re-design and improve the current diagnostic approaches, increasing the number of sequences used for their design and subsequently expanding the assays targets.

6.6 Future Work

An important outcome of this project is identifying the need to improve data management in the future. The vast amount of data generated creates a huge issue when having to manage quick sequencing run turnovers in a resource-limited setting. The metagenomic approach requires as much sequencing depth as possible so the runs are left for 24hr-48hr, easily producing 200-300 gigabytes of data, which need to be basecalled, transferred and backed-up. The biggest problem encountered was the slow speed of data transfers due to the huge number of files generated per run. The folder structure along with the file system of the nanopore data produced created significant handling difficulties. However, this is something that will be improved and easier to manage in the future, as the technology improves and the data generated is output in an easier to handle format. Currently major advancements have already been made since 2018, with a much simpler folder architecture and data generated in batches of 4,000 reads per file instead of one read per file, massively reducing data transfer times. Additionally, an important development is the MinIT, a preconfigured, small-footprint computer that supports MinION runs as it allows for the integration of live basecalling to the sequencing experiment. The ability to live basecall and the development of the flip-flop basecaller which now runs on GPU, provides faster data analysis and higher accuracy. Live basecalling and increased accuracy allow for improvements in the data analysis pipeline, with faster generation of fastq files and higher per read accuracy slowly moving towards making real-time analysis a reality. The improved per read accuracy leads to more accurate consensus generation without the need to use the raw fast5 files to polish the consensus sequence, eliminating the need to transfer the fast5 files and making data handling easier. Future improvements will streamline and optimise protocols for ease of use and will identify aspects of the current pipeline, both laboratory and analysis, which can be designed to be less time-consuming and more user-friendly. Developing a real-time automated pipeline will increase the data analysis efficiency and increase the ease of use. This will enable the completion of easy-to-follow protocols and standard operating procedures allowing for standardisation and training possibilities, particularly in-country. Capacity building is an important element, particularly when research such as in Chapter 5 takes place. During the 2018 outbreak the sequencing efforts were intensified and instead of the original pilot study, activities were increased and focused towards outbreak response. This led to limited opportunities to invest in developing standard operating procedures to allow for training and capacity building as all efforts were focused on outbreak support. Corresponding capacities must be

built in Nigeria and other African countries, allowing for the training of laboratory staff locally to enable sequencing expertise locally and allow for sequencing experiments to take place independently. This not only allows for capacity building and expanding expertise in-country but will also provide the opportunity for year-round surveillance to enhance our understanding of the virus epidemiology.

6.7 Conclusions

It is evident that mobile sequencing has an important role to play in disease management and response during outbreaks, epidemics and ultimately in ongoing surveillance. Mobile sequencing establishment in the heart of an unfolding viral outbreak or in locations with significant risk of a probable outbreak will have a major role in future case management and public health responses. Mobile sequencing should be standardly incorporated in support of response efforts of such future incidents, allowing for prompt implementation of genomic surveillance. Targeted methods will continue to provide large scale, high depth analysis of clonal outbreaks, nevertheless this thesis clearly highlights the role metagenomic methods can play in surveillance studies, outbreak of diverse pathogens and rapid response to novel/unexpected viral pathogens. Finally, it is apparent that sequencing efforts and the data generated from real-time mobile sequencing should be communicated in as near to real-time as possible to maximise the benefits of the approach.

Supplementary material

Figure S1 Pileup script

```
# run: python scriptname.py filename(without_bam_extention) fastafile(ref)
# filename = sys.argv[1]
# fastafile = sys.argv[2]

filename = sys.argv[1]
reference = sys.argv[2]
csvname = sys.argv[3]
fastaname = sys.argv[4]

header = ['Position', 'A', 'C', 'G', 'T', 'Insertions', 'Deletions', 'Consensus']
bamfile = pysam.AlignmentFile('{}.bam'.format(filename))
results = []
consensus = []
for record in pysamstats.stat_variation(bamfile,
                                        fafile= sys.argv[2]):
    rec =
[record['pos'],record['A'],record['C'],record['G'],record['T'],record['insertions'],record['deletions']]
    if rec[1]+rec[2]+rec[3]+rec[4]:
        percentage = float(max(rec[1:5])) / (rec[1]+rec[2]+rec[3]+rec[4])
    else:
        percentage = 0
        #print(record['pos'])
    ind = rec.index(max(rec[1:5]))
    if ind==1:
        found='A'
    elif ind==2:
        found='C'
    elif ind==3:
        found='G'
    elif ind==4:
        found='T'
    else:
        found='N'
    if max(rec[1:5])>19 and percentage>=0.7:
        rec.append(found)
        consensus.append(found)
    else:
        rec.append('N')
        consensus.append('N')
    results.append(rec)
    seq = ''.join(consensus) # create a string of GTACN

with open('{}.csv'.format(csvname), 'w', newline='') as csvfile:
    writer = csv.writer(csvname, delimiter=',')
    writer.writerow(i for i in header)
    writer.writerows(results)
with open('{}.fasta'.format(fastaname), 'w') as fastafile:
    fastafile.write('>{}\n'.format(filename))
    fastafile.write(seq)
```

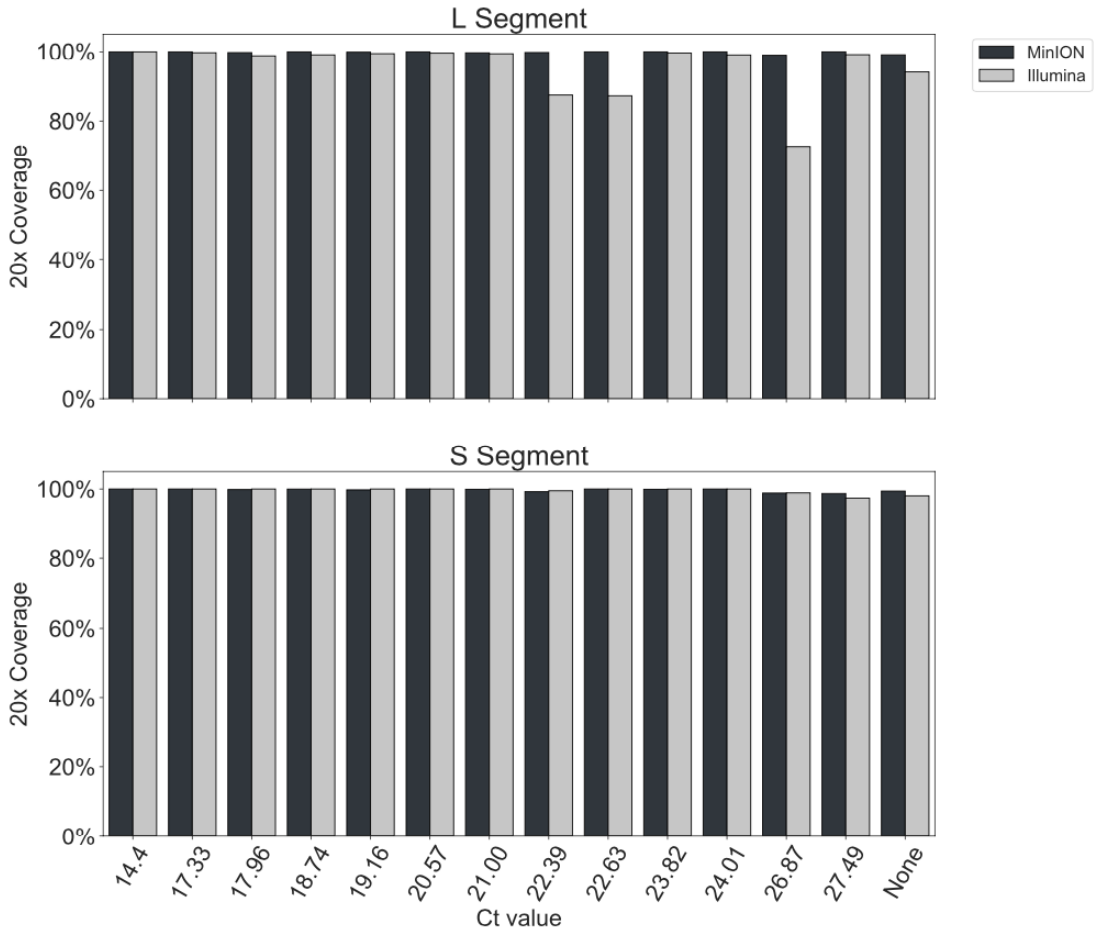


Figure S2 Proportion of reference genome recovered at minimum 20-fold coverage for each sample and both platforms (n = 14 samples).

Table S1. Description of samples positive for LASV by Altona real-time reverse transcription-PCR and by Nikisins real-time reverse transcription-PCR with corresponding percentage of MinION and illumina reads mapping to the L segment along with percentage of 20x genome coverage (n = 14 samples).

Sample Information			MinION		MiSeq	
Sample ID	Altona Ct Value	Nikisins Ct Value	L Segment % reads mapping	L Segment % 20x coverage	L Segment % reads mapping	L Segment % 20x coverage
ISTH-2018-073	14.4	21.18	15.65	100.00	15.2	100.00
ISTH-2018-119	17.33	16.36	10.04	100.00	9.6	99.74
ISTH-2018-126	17.96	20.5	6.86	99.79	6.24	98.84
ISTH-2018-066	18.74	26.73	20.66	100.00	14.4	99.13
ISTH-2018-014	19.16	26.42	1.67	99.93	3.13	99.48
ISTH-2018-131	20.57	25.68	1.34	100.00	0.96	99.64
ISTH-2018-075	22.39	29.74	1.35	99.82	0.05	87.43
ISTH-2018-072	22.63	34.3	8.44	100.00	0.13	87.20
ISTH-2018-021	23.82	31.76	1.56	100.00	1.21	99.67
ISTH-2018-013	24.01	26.02	1.78	100.00	0.77	99.10
ISTH-2018-074	26.87	32.9	0.43	99.02	0.05	72.69
ISTH-2018-001	27.49	23.75	4.57	100.00	2.58	99.18
ISTH-2018-115	21	26.55	1.35	99.72	1.41	99.44
ISTH-2018-036	NA	28.44	1.66	99.15	0.16	94.02

Table S2. Description of samples positive for LASV by Altona real-time reverse transcription-PCR and by Nikisins real-time reverse transcription-PCR with corresponding percentage of MinION and illumina reads mapping to the S segment along with percentage of 20x genome coverage (n = 14 samples).

Sample Information			MinION		MiSeq	
Sample ID	Altona Ct Value	Nikisins Ct Value	S Segment % reads mapping	S Segment % 20x coverage	S Segment % reads mapping	S Segment % 20x coverage
ISTH-2018-073	14.4	21.18	11.09	100.00	10.97	100.00
ISTH-2018-119	17.33	16.36	14.28	100.00	19.89	100.00
ISTH-2018-126	17.96	20.5	9.44	99.85	14.2	100.00
ISTH-2018-066	18.74	26.73	15.8	99.97	18.23	100.00
ISTH-2018-014	19.16	26.42	2.99	99.77	7	100.00
ISTH-2018-131	20.57	25.68	0.81	100.00	0.72	100.00
ISTH-2018-075	22.39	29.74	1.1	99.24	0.07	99.50
ISTH-2018-072	22.63	34.3	8.04	100.00	0.09	100.00
ISTH-2018-021	23.82	31.76	1.41	99.91	1.73	100.00
ISTH-2018-013	24.01	26.02	1.36	100.00	0.43	100.00
ISTH-2018-074	26.87	32.9	0.26	98.84	0.05	98.90
ISTH-2018-001	27.49	23.75	2.66	98.71	0.66	97.39
ISTH-2018-115	21	26.55	1.74	99.91	2.25	100.00
ISTH-2018-036	NA	28.44	0.72	99.41	0.07	98.05

Table S3. Comparison between Nanopore and Illumina consensus sequences.

Sample ID ISTH-2018-	Altona Ct Value	Nikisins Ct Value	Segment Length	Total Disagreements	In Coding Regions	Positions (Illumina > Nanopore, non-coding/coding)
L Segment						
073	14.4	21.18	7260	4	3	5792G>T, 5793A>G, 5795G>A, 7258C>T
119	17.33	16.36	7258	2	0	404C>T, 407C>T
126	17.96	20.5	7245	0	0	
066	18.74	26.73	7135	1	1	274G>T
014	19.16	26.42	7245	1	0	445G>A
131	20.57	25.68	7238	0	0	
115	21.00	26.55	7183	3	3	727C>T, 728T>C, 731C>T
075	22.39	29.74	7256	0	0	
072	22.63	34.3	7250	0	0	
021	23.82	31.76	7237	0	0	
013	24.01	26.02	7230	0	0	
074	26.87	32.9	7238	0	0	
001	27.49	23.75	7196	2	0	2A>T, 416T>G
036	None	28.44	7261	1	0	7209T>C
S Segment						
073	14.4	21.18	3406	1	0	3422C>G
119	17.33	16.36	3403	2	2	596C>A, 597A>C
126	17.96	20.5	3407	1	0	2C>G
066	18.74	26.73	3393	0	0	
014	19.16	26.42	3407	1	0	1587T>G
131	20.57	25.68	3389	0	0	
115	21.00	26.55	3387	0	0	
075	22.39	29.74	3412	0	0	
072	22.63	34.3	3398	6	0	1554C>G, 1568C>G, 1569G>A, 1570C>A, 1572G>A, 1573A>G
021	23.82	31.76	3385	0	0	
013	24.01	26.02	3367	0	0	
074	26.87	32.9	3367	0	0	
001	27.49	23.75	3490	0	0	
036	None	28.44	3387	0	0	

NCDC full report. Early Results of Lassa Virus Sequencing & Implications for Current Outbreak Response in Nigeria

12 March, 2018 | Abuja – EARLY RESULTS OF LASSA VIRUS SEQUENCING AND IMPLICATIONS FOR CURRENT OUTBREAK RESPONSE IN NIGERIA

The Nigeria Centre for Disease Control continues to lead a multi-agency, multi-partner response to the on-going Lassa fever outbreak in Nigeria. The response is now gathering steam with an escalation of efforts to control spread in all the foci of infections across the country. A critical aspect of outbreak control is research, to rapidly understand the drivers of infection and opportunities for control.

A relatively new tool available to epidemiologists and researchers is the use of whole genome sequencing during outbreaks. Real-time sequencing of viruses can reveal crucial details that contribute to the understanding and the control of infectious disease outbreaks. It is now possible to resolve chains of transmission to a level of detail otherwise unachievable using traditional methods.

Researchers at the Irrua Specialist Teaching Hospital, Edo State, in collaboration with partners from the Bernhard-Nocht Institute for Tropical Medicine, Germany, and others (see below) have deployed real-time sequencing of Lassa fever viruses from the on-going outbreak to learn key lessons to influence control efforts. The real-time availability of sequencing information for the current 2018 Lassa fever viruses will support the response to the on-going Lassa fever outbreak.

On the 10th March 2018, the analysis of a first set of seven Lassa virus draft sequences derived from the blood of seven laboratory-confirmed Lassa fever patients from Edo, Ondo, Ebonyi, and Imo States was completed. In addition, 83 unpublished sequences obtained during previous years from various parts of Nigeria were made available for comparison with the sequences of the 2018 outbreak. From the analysis of these results, we can draw the following preliminary conclusions:

• No new virus lineages have so far been detected, meaning that the circulating viruses appear to be very similar to the viruses from previous years;

• Lassa fever cases appear to be mostly caused by viruses that are not epidemiologically linked;

• The viruses circulating in 2018 appear to originate from the pool of lineages and strains known to be circulating in Nigeria and are consistent with previous outbreaks;

• The most likely route of transmission continues to be spill over of viruses from the rodent reservoir to humans rather than extensive human-to-human transmission.

The most important question being investigated at the moment is what has caused an outbreak of this magnitude, at this time? One of the possible answers to this question is the emergence of a new Lassa virus lineage or strain with increased virulence or transmissibility. Evidence from this work, although limited to seven viruses at the moment suggests that this is unlikely to be the case.

Over the last year, the Nigeria Centre for Disease Control has improved its disease surveillance, detection and response system. This includes the operationalisation of the National Reference Laboratory in Abuja, to add to existing diagnostic capacity, with its capacity to test for Lassa fever in ISTH and Lagos University Teaching Hospital (LUTH). According to Chief Executive Office, Dr. Chikwe Ihekweazu, “As the system continues to be strengthened, there is likely to be an

increase in the number of cases detected and reported. Sequencing provides an exciting additional tool for the outbreak response. As we learn more, we become better equipped to respond."

To support the public health response as well as the development and evaluation of Lassa fever diagnostics or therapeutics, the researchers are sharing the sequences via the website virological.org.

The NCDC would like to thank the following institutions and partners for their continued support to the Government of Nigeria:

• Irrua Specialist Teaching Hospital (ISTH), Institute of Lassa Fever Research and Control (ILFRC), Irrua, Edo State, Nigeria (Prof. Sylvanus Okogbenin, Dr. Ephraim Ogbaini, Dr. Adomeh Donatus)

• Bernhard Nocht Institute for Tropical Medicine (BNITM), Hamburg, Germany (Prof. Stephan Günther, Liana Kafetzopoulou, Dr. Deborah Ehichioya, Dr. Meike Pahlmann, Dr. Lisa Oestereich, Dr. Sophie Duraffour, Anke Thielebein, Julia Hinzmann)

• Public Health England (PHE), Salisbury, United Kingdom (Prof. Miles Carroll, Liana Kafetzopoulou, Dr. Steven Pullan, Dr. Richard Vipond, Dr. Roger Hewson)

• KU Leuven, Leuven, Belgium (Dr. Philippe Lemey)

• University of Liverpool, United Kingdom (Dr. Julian Hiscox)

• World Health Organisation (WHO)

Contacts

NCDC Toll-free Number: 0800-970000-10

SMS: 08099555577

WhatsApp: 07087110839

Twitter/Facebook: @NCDCgov

Signed:

Dr. Chikwe Ihekweazu,

CEO, Nigeria Centre for Disease Control

References

References

1. International Committee on Taxonomy of Viruses (ICTV). *International Committee on Taxonomy of Viruses (ICTV)*, (available at https://talk.ictvonline.org/taxonomy/w/ictv-taxonomy).

2. D. Baltimore, Expression of animal virus genomes. *Bacteriol. Rev.* 35, 235–241 (1971).

3. B. N. Fields, *Fields Virology* (Lippincott Williams & Wilkins, 2013).

4. M. E. J. Woolhouse, L. Brierley, Epidemiological characteristics of human-infective RNA viruses. *Sci Data*. 5, 180017 (2018).

5. M. Woolhouse, Sources of human viruses. *Science*. 362, 524–525 (2018).

6. J. Ortín, J. Martín-Benito, The RNA synthesis machinery of negative-stranded RNA viruses. *Virology*. 479-480, 532–544 (2015).

7. F. Begum, C. L. Wisseman Jr, J. Casals, TICK-BORNE VIRUSES OF WEST PAKISTAN: II. HAZARA VIRUS, A NEW AGENT ISOLATED FROM IXODES REDIKORZEVI TICKS FROM THE KAGHAN VALLEY, W. PAKISTAN. *Am. J. Epidemiol.* 92, 192–194 (1970).

8. O. Ergonul, C. A. Whitehouse, *Crimean-Congo Hemorrhagic Fever: A Global Perspective* (Springer Science & Business Media, 2007).

9. M. A. Darwish, H. Hoogstraal, T. J. Roberts, R. Ghazi, T. Amer, A sero-epidemiological survey for Bunyaviridae and certain other arboviruses in Pakistan. *Transactions of the Royal Society of Tropical Medicine and Hygiene*. 77, 446–450 (1983).

10. R. A. Surtees, thesis, University of Leeds (2014).

11. Y. Matsumoto, K. Ohta, D. Kolakofsky, M. Nishio, A Minigenome Study of Hazara Nairovirus Genomic Promoters. *J. Virol.* 93 (2019).

12. S. D. Dowall, S. Findlay-Wilson, E. Rayner, et al., Hazara virus infection is lethal for adult type I interferon receptor-knockout mice and may act as a surrogate for infection with the human-pathogenic Crimean-Congo hemorrhagic fever virus. *J. Gen. Virol.* 93, 560–564 (2012).

13. International Committee on Taxonomy of Viruses (ICTV). *International Committee on Taxonomy of Viruses (ICTV)*, (available at https://talk.ictvonline.org/).

14. S. H. Ishak, L. H. Yaacob, A. Ishak, Severe Dengue with Hemophagocytosis Syndrome. *Malays Fam Physician*. 15, 47–49 (2020).

15. B. E. E. Martina, P. Koraka, A. D. M. E. Osterhaus, Dengue virus pathogenesis: an integrated view. *Clin. Microbiol. Rev.* 22, 564–581 (2009).

16. T. Vos, C. Allen, M. Arora, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–

2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet*. 388, 1545–1602 (2016).

17. Dengue and severe dengue, (available at https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue).

18. J. Patterson, M. Sammon, M. Garg, Dengue, Zika and Chikungunya: Emerging Arboviruses in the New World. *West. J. Emerg. Med.* 17, 671–679 (2016).

19. S. Payne, in *Viruses*, S. Payne, Ed. (Academic Press),129–139 (2017).

20. S. M. Paranjape, E. Harris, Control of dengue virus translation and replication. *Curr. Top. Microbiol. Immunol.* 338, 15–34 (2010).

21. M. R. Holbrook, *Advances in Flavivirus Research* (MDPI, 2018).

22. S.-D. Thiberville, N. Moyen, L. Dupuis-Maguiraga, et al., Chikungunya fever: Epidemiology, clinical syndrome, pathogenesis and therapy. *Antiviral Res.* 99, 345–370 (2013).

23. WHO | WHO publishes list of top emerging diseases likely to cause major epidemics (2017) (available at http://www.who.int/medicines/ebola-treatment/WHO-list-of-top-emerging-diseases/en/).

24. S. C. Weaver, M. Lecuit, Chikungunya virus and the global spread of a mosquito-borne disease. *N. Engl. J. Med.* 372, 1231–1239 (2015).

25. T. Couderc, M. Lecuit, Chikungunya virus pathogenesis: From bedside to bench. *Antiviral Res.* 121, 120–131 (2015).

26. F. J. Burt, W. Chen, J. J. Miner, et al., Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen. *Lancet Infect. Dis.* 17, e107–e117 (2017).

27. D. Mavalankar et al, Increased Mortality Rate Associated with Chikungunya Epidemic, Ahmedabad, India - Volume 14, Number 3—March 2008 - Emerging Infectious Disease journal - CDC, 14.3, 412 (2008)

28. S. Payne, in *Viruses*, S. Payne, Ed. (Academic Press, 141–148 (2017)

29. M. Solignat, B. Gay, S. Higgs, L. Briant, C. Devaux, Replication cycle of chikungunya: a re-emerging arbovirus. *Virology*. 393, 183–197 (2009).

30. F. J. Burt, M. S. Rolph, N. E. Rulli, S. Mahalingam, M. T. Heise, Chikungunya: a re-emerging virus. *Lancet*. 379, 662–671 (2012).

31. C. P. Simmons, J. J. Farrar, van V. C. Nguyen, B. Wills, Dengue. *N. Engl. J. Med.* 366, 1423–1432 (2012).

32. L. Furuya-Kanamori, S. Liang, G. Milinovich, et al., Co-distribution and co-infection of chikungunya and dengue viruses. *BMC Infect. Dis.* 16, 84 (2016).

33. M. Perera-Lecoin, N. Luplertlop, P. Surasombatpattana, et al., in *Current Topics in Chikungunya*, A. J. Rodriguez-Morales, Ed. (InTech, 2016).

34. R. Omarjee, C. M. Prat, O. Flusin, et al., Importance of case definition to monitor

ongoing outbreak of chikungunya virus on a background of actively circulating dengue virus, St Martin, December 2013 to January 2014. *Eurosurveillance*. 19, 20753 (2014).

35. C. A. A. Brito, F. Azevedo, M. T. Cordeiro, E. T. A. Marques Jr, R. F. O. Franca, Central and peripheral nervous system involvement caused by Zika and chikungunya coinfection. *PLoS Negl. Trop. Dis.* 11, e0005583 (2017).

36. A. Wilder-Smith, D. J. Gubler, S. C. Weaver, T. P. Monath, D. L. Heymann, T. W. Scott, Epidemic arboviral diseases: priorities for research and public health. *Lancet Infect. Dis.* 17, e101–e106 (2017).

37. Nigeria Centre for Disease Control, (available at https://ncdc.gov.ng/diseases/factsheet/30).

38. Lassa Fever. *WHO | Regional Office for Africa*, (available at https://www.afro.who.int/health-topics/lassa-fever).

39. D. A. Asogun, D. I. Adomeh, J. Ehimuan, et al., Molecular diagnostics for lassa fever at Irrua specialist teaching hospital, Nigeria: lessons learnt from two years of laboratory operation. *PLoS Negl. Trop. Dis.* 6, e1839 (2012).

40. J. D. Frame, J. M. Baldwin Jr, D. J. Gocke, J. M. Troup, Lassa fever, a new virus disease of man from West Africa. I. Clinical description and pathological findings. *Am. J. Trop. Med. Hyg.* 19, 670–676 (1970).

41. M. D. Bowen, P. E. Rollin, T. G. Ksiazek, et al., Genetic diversity among Lassa virus strains. *J. Virol.* 74, 6992–7004 (2000).

42. E. Fichet-Calvet, D. J. Rogers, Risk maps of Lassa fever in West Africa. *PLoS Negl. Trop. Dis.* 3, e388 (2009).

43. S. Atkin, S. Anaraki, P. Gothard, et al., The first case of Lassa fever imported from Mali to the United Kingdom, February 2009. *Euro Surveill.* 14 (2009).

44. S. Günther, P. Emmerich, T. Laue, et al., Imported lassa fever in Germany: molecular characterization of a new lassa virus strain. *Emerg. Infect. Dis.* 6, 466–476 (2000).

45. Lassa fever. *World Health Organization*, (available at http://www.who.int/news-room/fact-sheets/detail/lassa-fever).

46. J. B. McCormick, S. P. Fisher-Hoch, Lassa fever. *Curr. Top. Microbiol. Immunol.* 262, 75–109 (2002).

47. G. Lo Iacono, A. A. Cunningham, E. Fichet-Calvet, et al., Using modelling to disentangle the relative contributions of zoonotic and anthroponotic transmission: the case of lassa fever. *PLoS Negl. Trop. Dis.* 9, e3398 (2015).

48. K. G. Andersen, B. J. Shapiro, C. B. Matranga, et al., Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. *Cell*. 162, 738–750 (2015).

49. S. Payne, in *Viruses*, S. Payne, Ed. (Academic Press), 215–218 (2017).

50. S. Günther, O. Lenz, Lassa virus. *Crit. Rev. Clin. Lab. Sci.* 41, 339–390 (2004).

51. R. Klitting, S. B. Mehta, J. U. Oguzie, et al., Lassa Virus Genetics. *Curr. Top. Microbiol. Immunol.* (2020).

52. *Molecular Virology of Human Pathogenic Viruses* (Elsevier, 2017).

53. Arenaviridae - Negative Sense RNA Viruses - Negative Sense RNA Viruses (2011) - International Committee on Taxonomy of Viruses (ICTV). *International Committee on Taxonomy of Viruses (ICTV)*, (available at https://talk.ictvonline.org/ictv-reports/ictv_9th_report/negative-sense-rna-viruses-2011/w/negrna_viruses/203/arenaviridae).

54. M. Lehmann, M. Pahlmann, H. Jérôme, C. Busch, M. Lelke, S. Günther, Role of the C terminus of Lassa virus L protein in viral mRNA synthesis. *J. Virol.* 88, 8713–8717 (2014).

55. N. E. Yun, D. H. Walker, Pathogenesis of Lassa fever. *Viruses*. 4, 2031–2048 (2012).

56. X. Qi, S. Lan, W. Wang, et al., Cap binding and immune evasion revealed by Lassa nucleoprotein structure. *Nature*. 468, 779–783 (2010).

57. M. Lelke, L. Brunotte, C. Busch, S. Günther, An N-terminal region of Lassa virus L protein plays a critical role in transcription but not replication of the virus genome. *J. Virol.* 84, 1934–1944 (2010).

58. B. Morin, B. Coutard, M. Lelke, et al., The N-terminal domain of the arenavirus L protein is an RNA endonuclease essential in mRNA transcription. *PLoS Pathog.* 6, e1001038 (2010).

59. M. S. Salvato, E. M. Shimomaye,The completed sequence of lymphocytic choriomeningitis virus reveals a unique RNA structure and a gene for a zinc finger protein. *Virology*. 173, 1–10 (1989).

60. M. Kiening, F. Weber, D. Frishman, Conserved RNA structures in the intergenic regions of ambisense viruses. *Sci. Rep.* 7, 16625 (2017).

61. D. J. Gubler, G. G. Clark, Dengue/dengue hemorrhagic fever: the emergence of a global health problem. *Emerg. Infect. Dis.* 1, 55–57 (1995).

62. S. S. Morse, Factors in the Emergence of Infectious Diseases. *Emerging Infectious Diseases*. 1, 7–15 (1995).

63. M. E. J. Woolhouse, D. T. Haydon, R. Antia, Emerging pathogens: the epidemiology and evolution of species jumps. *Trends in Ecology & Evolution*. 20, 238–244 (2015).

64. History of Ebola Virus Disease Error processing SSI file (2019), (available at https://www.cdc.gov/vhf/ebola/history/summaries.html).

65. National Institutes of Health (US), Biological Sciences Curriculum Study, in *NIH Curriculum Supplement Series [Internet]* (National Institutes of Health (US), 2007).

66. WHO | The West African situation (2016) (available at http://www.who.int/csr/disease/yellowfev/west-africa-situation/en/).

67. Website, (available at https://www.who.int/emergencies/diseases/en/).

68. M. M. Akiner, B. Demirci, G. Babuadze, V. Robert, F. Schaffner, Spread of the Invasive Mosquitoes Aedes aegypti and Aedes albopictus in the Black Sea Region Increases Risk of Chikungunya, Dengue, and Zika Outbreaks in Europe. *PLoS Negl. Trop. Dis.* 10, e0004664 (2016).

69. S. V. Mayer, R. B. Tesh, N. Vasilakis, The emergence of arthropod-borne viral diseases: A global prospective on dengue, chikungunya and zika fevers. *Acta Trop.* 166, 155–163 (2017).

70. S. L. Smits, A. D. Osterhaus, Virus discovery: one step beyond. *Curr. Opin. Virol.* 3, e1–e6 (2013).

71. S. S. Morse, J. A. K. Mazet, M. Woolhouse, et al., Prediction and prevention of the next pandemic zoonosis. *Lancet*. 380, 1956–1965 (2012).

72. S. Bedhomme, J. Hillung, S. F. Elena, Emerging viruses: why they are not jacks of all trades? *Current Opinion in Virology*. 10, 1–6 (2015).

73. N. H. Ogden, P. AbdelMalik, J. Pulliam, Emerging infectious diseases: prediction and detection. *Can. Commun. Dis. Rep.* 43, 206–211 (2017).

74. B. L. Haagmans, A. C. Andeweg, A. D. M. E. Osterhaus, The application of genomics to emerging zoonotic viral diseases. *PLoS Pathog.* 5, e1000557 (2009).

75. Forum on Microbial Threats, Board on Global Health, Institute of Medicine, *Emerging Viral Diseases: The One Health Connection: Workshop Summary* (National Academies Press (US), Washington (DC), 2015).

76. M. E. J. Woolhouse, S. Gowtage-Sequeria, Host range and emerging and reemerging pathogens. *Emerg. Infect. Dis.* 11, 1842–1847 (2005).

77. K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, P. Daszak, Global trends in emerging infectious diseases. *Nature*. 451, 990–993 (2018).

78. P. G. Coleman, E. M. Fèvre, S. Cleaveland, Estimating the public health impact of rabies. *Emerg. Infect. Dis.* 10, 140–142 (2004).

79. C. R. Fisher, D. G. Streicker, M. J. Schnell, The spread and evolution of rabies virus: conquering new frontiers. *Nat. Rev. Microbiol.* 16, 241–255 (2018).

80. E. Lecompte, E. Fichet-Calvet, S. Daffis, et al., Mastomys natalensis and Lassa fever, West Africa. *Emerg. Infect. Dis.* 12, 1971–1974 (2006).

81. A. Olayemi, A. Obadare, A. Oyeyiola, et al., Arenavirus Diversity and Phylogeography of Mastomys natalensis Rodents, Nigeria. *Emerg. Infect. Dis.* 22, 694–697 (2016).

82. L.-F. Wang, L. -F. Wang, G. Crameri, Emerging zoonotic viral diseases. *Revue Scientifique et Technique de l'OIE*. 33, 569–581 (2014).

83. A. D. T. Barrett, West Nile in Europe: an increasing public health problem. *J. Travel Med.* 25 (2018).

84. N. Vonesch, A. Binazzi, M. Bonafede, P. Melis, A. Ruggieri, S. Iavicoli, P. Tomao, Emerging zoonotic viral infections of occupational health importance. *Pathog. Dis.* 77 (2019).

85. G. Kuno, G.-J. J. Chang, Biological transmission of arboviruses: reexamination of and new insights into components, mechanisms, and unique traits as well as their evolutionary trends. *Clin. Microbiol. Rev.* 18, 608–637 (2005).

86. A. Papa, Emerging arboviral human diseases in Southern Europe. *J. Med. Virol.* 89, 1315–1322 (2017).

87. S. C. Weaver, W. K. Reisen, Present and future arboviral threats. *Antiviral Res.* 85, 328–345 (2010).

88. E. Gould, J. Pettersson, S. Higgs, R. Charrel, X. de Lamballerie, Emerging arboviruses: Why today? *One Health*. 4, 1–13 (2017).

89. N. Cleton, M. Koopmans, J. Reimerink, G.-J. Godeke, C. Reusken, Come fly with me: review of clinically important arboviruses for global travelers. *J. Clin. Virol.* 55, 191–203 (2012).

90. B. McCloskey, O. Dar, A. Zumla, D. L. Heymann, Emerging infectious diseases and pandemic potential: status quo and reducing risk of global spread. *The Lancet Infectious Diseases*. 14, 1001–1010 (2014).

91. S. C. Weaver, C. Charlier, N. Vasilakis, M. Lecuit, Zika, Chikungunya, and Other Emerging Vector-Borne Viral Diseases. *Annu. Rev. Med.* 69, 395–408 (2018).

92. O. G. Pérez, *Astrocytes: Structure, Functions and Role in Disease* (Nova Science Publishers, Incorporated, 2012).

93. P. G. E. Kennedy, VIRAL ENCEPHALITIS: CAUSES, DIFFERENTIAL DIAGNOSIS, AND MANAGEMENT. *Journal of Neurology, Neurosurgery & Psychiatry*. 75, 10i–15 (2004).

94. M. Ellul, T. Solomon, Acute encephalitis – diagnosis and management. *Clinical Medicine*. 18, 155–159 (2018).

95. Factsheet about tick-borne encephalitis (TBE). *European Centre for Disease Prevention and Control*, (available at http://ecdc.europa.eu/en/tick-borne-encephalitis/facts/factsheet).

96. WHO | Haemorrhagic fevers, Viral (2018) (available at http://www.who.int/topics/haemorrhagic_fevers_viral/en/).

97. A. Brinkmann, K. Ergünay, A. Radonić, Z. Kocak Tufan, C. Domingo, A. Nitsche, Development and preliminary evaluation of a multiplexed amplification and next generation sequencing method for viral hemorrhagic fever diagnostics. *PLoS Negl. Trop. Dis.* 11, e0006075 (2017).

98. C. Drosten, B. M. Kümmerer, H. Schmitz, S. Günther, Molecular diagnostics of viral hemorrhagic fevers. *Antiviral Research*. 57, 61–87 (2003).

99. Lassa Fever. *WHO | Regional Office for Africa*, (available at https://www.afro.who.int/health-topics/lassa-fever).

100. A. L. Rasmussen, M. G. Katze, Genomic Signatures of Emerging Viruses: A New Era of Systems Epidemiology. *Cell Host Microbe*. 19, 611–618 (2016).

101. E. C. Chen, S. A. Miller, J. L. DeRisi, C. Y. Chiu, Using a Pan-Viral Microarray Assay (Virochip) to Screen Clinical Samples for Viral Pathogens. *Journal of Visualized Experiments* (2011).

102. C. J. Houldcroft, M. A. Beale, J. Breuer, Clinical and biological insights from viral genome sequencing. *Nat. Rev. Microbiol.* 15, 183–192 (2017).

103. N. D. Grubaugh, J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, K. G. Andersen, Tracking virus outbreaks in the twenty-first century. *Nat Microbiol*. 4, 10–19 (2019).

104. A. Mayer, D. Ivanowski, M. W. Beijerinck, E. Baur, in *Early Papers on Tobacco Mosaic and Infectious Variegation* (The American Phytopathological Society, 1942), *Phytopathological Classics Series*, 9–62 (1942).

105. L. Bos, Beijerinck's work on tobacco mosaic virus: historical context and legacy. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 675–685 (1999).

106. F. Brown, The history of research in foot-and-mouth disease. *Virus Res.* 91, 3–7 (2003).

107. A. D. Barrett, T. P. Monath, Epidemiology and ecology of yellow fever virus. *Adv. Virus Res.* 61, 291–315 (2003).

108. C. S. Bryan, S. W. Moss, R. J. Kahn, Yellow fever in the Americas. *Infect. Dis. Clin. North Am.*, 18.2, 275-92 (2004)

109. J. A. del Regato, James Carroll: a biography. *Ann. Diagn. Pathol.* 2, 335–349 (1998).

110. D. M. Knipe, P. M. Howley, *Fields' Virology* (Lippincott Williams & Wilkins, 2007).

111. C. B. Matranga, K. G. Andersen, S. Winnicki, et al., Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* 15, 519 (2014).

112. L. Barzon, E. Lavezzo, G. Costanzi, E. Franchin, S. Toppo, G. Palù, Next-generation sequencing technologies in diagnostic virology. *Journal of Clinical Virology*. 58, 346–350 (2013).

113. C. Chiu, Clinical metagenomics for diagnosis and discovery of viral pathogens. *Nature reviews, Genetics* 20.6, 341 (2012)

114. R. W. Peeling, H. Artsob, J. L. Pelegrino, et al., Evaluation of diagnostic tests: dengue. *Nat. Rev. Microbiol.* 8, S30–8 (2010).

115. S. Yang, R. E. Rothman, PCR-based diagnostics for infectious diseases: uses, limitations, and future applications in acute-care settings. *Lancet Infect. Dis.* 4, 337–348 (2004).

116. P. Martin, K. B. Laupland, E. H. Frost, L. Valiquette, Laboratory diagnosis of Ebola virus disease. *Intensive Care Medicine*. 41, 895–898 (2015).

117. S. A. Bustin, R. Mueller, Real-time reverse transcription PCR (qRT-PCR) and its potential use in clinical diagnosis. *Clinical Science*. 109, 365–379 (2005).

118. G. A. Storch, Diagnostic Virology. *Clin. Infect. Dis.* 31, 739–751 (2000).

119. D. A. Muller, A. C. I. Depelsenaire, P. R. Young, Clinical and Laboratory Diagnosis of Dengue Virus Infection. *The Journal of Infectious Diseases*. 215, S89–S95 (2017).

120. C. W. H. Michael T. Osterholm, Epidemiologic Principles. *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*, 146 (2015).

121. Principles of Epidemiology: Lesson 5, Section 2|Self-Study Course SS1978|CDC (2020), (available at https://www.cdc.gov/csels/dsepd/ss1978/lesson5/section2.html).

122. T. A. Brown, *Molecular Phylogenetics* (Wiley-Liss, 2002).

123. N. Papavero, Essays on the History of Neotropical Dipterology V. 2 (1973).

124. G. Dudas, T. Bedford, The ability of single genes vs full genomes to resolve time and space in outbreak analysis. *BMC Evolutionary Biology*. 19 (2019).

125. WHO | Surveillance (2017) (available at https://www.who.int/ihr/surveillance/en/).

126. A. D. M. E. Osterhaus, S. L. Smits, in *Genomic and Personalized Medicine (Second Edition)*, G. S. Ginsburg, H. F. Willard, Eds. (Academic Press), 1142–1154 (2013).

127. A. J. Drummond, O. G. Pybus, A. Rambaut, R. Forsberg, A. G. Rodrigo, Measurably evolving populations. *Trends Ecol. Evol.* 18, 481–488 (2003).

128. R. Biek, O. G. Pybus, J. O. Lloyd-Smith, X. Didelot, Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* 30, 306–313 (2015).

129. WHO | SARS (Severe Acute Respiratory Syndrome) (2012) (available at https://www.who.int/ith/diseases/sars/en/).

130. P. Jsm, J. S. M. Peiris, S. T. Lai, L. L. M. Poon, G. Yakan, L. Y. C. Yam, W. Lim, Coronavirus as a Possible Cause of Severe Acute Respiratory Syndrome. *The Journal of Tepecik Education and Research Hospital*. 13, 55–56 (2003).

131. C. Drosten, S. Günther, W. Preiser, et al., Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *N. Engl. J. Med.* 348, 1967–1976 (2003).

132. T. G. Ksiazek, D. Erdman, C. S. Goldsmith, et al., SARS Working Group, A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* 348, 1953–1966 (2003).

133. CDC, 2009 H1N1 Flu Pandemic Timeline. *Centers for Disease Control and Prevention* (2020), (available at https://www.cdc.gov/flu/pandemic-resources/2009-pandemic-timeline.html).

134. C. Fraser, C. A. Donnelly, S. Cauchemez, et al., WHO Rapid Pandemic

Assessment Collaboration, Pandemic potential of a strain of influenza A (H1N1): early findings. *Science*. 324, 1557–1561 (2009).

135. A. Rambaut, E. Holmes, The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. *PLoS Curr.* 1, RRN1003 (2009).

136. G. J. D. Smith, D. Vijaykrishna, J. Bahl, et al., Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*. 459, 1122–1125 (2009).

137. Who Mers-Cov Research Group, State of Knowledge and Data Gaps of Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in Humans. *PLoS Curr.* 5 (2013).

138. B. L. Haagmans, S. H. S. Al Dhahiry, C. B. E. M. Reusken, et al., Middle East respiratory syndrome coronavirus in dromedary camels: an outbreak investigation. *Lancet Infect. Dis.* 14, 140–145 (2014).

139. E. I. Azhar, S. A. El-Kafrawy, S. A. Farraj, A. M. Hassan, M. S. Al-Saeed, A. M. Hashem, T. A. Madani, Evidence for camel-to-human transmission of MERS coronavirus. *N. Engl. J. Med.* 370, 2499–2505 (2014).

140. J. S. M. Sabir, T. T.-Y. Lam, M. M. M. Ahmed, L. Li, et al., Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science*. 351, 81–84 (2016).

141. E. C. Holmes, G. Dudas, A. Rambaut, K. G. Andersen, The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature*. 538, 193–200 (2016).

142. A. Arias, S. J. Watson, D. Asogun, et al., Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol*. 2, vew016 (2016).

143. S. Baize, D. Pannetier, L. Oestereich, et al., Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.* 371, 1418–1425 (2014).

144. S. K. Gire, A. Goba, K. G. Andersen, et al., Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*. 345, 1369–1372 (2014).

145. M. W. Carroll, D. A. Matthews, J. A. Hiscox, et al., Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. *Nature*. 524, 97–101 (2015).

146. G. Dudas, A. Rambaut, Phylogenetic Analysis of Guinea 2014 EBOV Ebolavirus Outbreak. *PLoS Curr.* 6 (2014).

147. J. Quick, N. J. Loman, S. Duraffour, et al., Real-time, portable genome sequencing for Ebola surveillance. *Nature*. 530, 228–232 (2016).

148. D. J. Park, G. Dudas, S. Wohl, et al., Ebola Virus Epidemiology, Transmission, and Evolution during Seven Months in Sierra Leone. *Cell*. 161, 1516–1526 (2015).

149. E. Simon-Loriere, O. Faye, O. Faye, et al., Distinct lineages of Ebola virus in

Guinea during the 2014 West African epidemic. *Nature*. 524, 102–104 (2015).

150. Y.-G. Tong, W.-F. Shi, D. Liu, et al., China Mobile Laboratory Testing Team in Sierra Leone, Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature*. 524, 93–96 (2015).

151. J. T. Ladner, M. R. Wiley, S. Mate, et al., Evolution and Spread of Ebola Virus in Liberia, 2014-2015. *Cell Host Microbe*. 18, 659–669 (2015).

152. T. Hoenen, A. Groseth, K. Rosenke, et al., Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. *Emerg. Infect. Dis.* 22, 331–334 (2016).

153. S. L. Smits, S. D. Pas, C. B. Reusken, et al., Genotypic anomaly in Ebola virus strains circulating in Magazine Wharf area, Freetown, Sierra Leone, 2015. *Euro Surveill.* 20 (2015).

154. WHO | Zika situation report (2016) (available at http://www.who.int/emergencies/zika-virus/situation-report/5-february-2016/en/).

155. About, (available at https://www.zibraproject.org/about/).

156. N. R. Faria, J. Quick, I. M. Claro, et al., Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature*. 546, 406–410 (2017).

157. N. R. Faria, R. do S. da S. Azevedo, M. U. G. Kraemer, et al., Zika virus in the Americas: Early epidemiological and genetic findings. *Science*. 352, 345–349 (2016).

158. H. C. Metsky, C. B. Matranga, S. Wohl, et al., Zika virus evolution and spread in the Americas. *Nature*. 546, 411–415 (2017).

159. N. D. Grubaugh, J. T. Ladner, M. U. G. Kraemer, et al., Genomic epidemiology reveals multiple introductions of Zika virus into the United States. *Nature*. 546, 401–405 (2017).

160. J. L. Gardy, N. J. Loman, Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.* 19, 9–20 (2018).

161. R. E. Franklin, R. G. Gosling, Molecular configuration in sodium thymonucleate. *Nature*. 171, 740–741 (1953).

162. J. D. Watson, F. H. C. Crick, Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*. 171, 737–738 (1953).

163. I. R. Lehman, S. B. Zimmerman, J. Adler, M. J. Bessman, E. S. Simms, A. Kornberg, ENZYMATIC SYNTHESIS OF DEOXYRIBONUCLEIC ACID. V. CHEMICAL COMPOSITION OF ENZYMATICALLY SYNTHESIZED DEOXYRIBONUCLEIC ACID. *Proc. Natl. Acad. Sci. U. S. A.* 44, 1191–1196 (1958).

164. J. Hurwitz, A. Bresler, R. Diringer, The enzymic incorporation of ribonucleotides into polyribonucleotides and the effect of DNA. *Biochem. Biophys. Res. Commun.* 3, 15–19 (1960).

165. R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill,

J. R. Penswick, A. Zamir, Structure of a Ribonucleic Acid. *Science*. 147, 1462–1465 (1965).

166. J. M. Heather, B. Chain, The sequence of sequencers: The history of sequencing DNA. *Genomics*. 107, 1–8 (2016).

167. (available at https://www.nobelprize.org/uploads/2018/06/holley-lecture.pdf).

168. F. Sanger, G. G. Brownlee, B. G. Barrell, A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* 13, 373–398 (1965).

169. W. Min Jou, G. Haegeman, M. Ysebaert, W. Fiers, Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*. 237, 82–88 (1972).

170. W. Fiers, R. Contreras, F. Duerinck, et al., Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*. 260, 500–507 (1976).

171. R. Wu, A. D. Kaiser, Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.* 35, 523–537 (1968).

172. R. Wu, Nucleotide Sequence Analysis of DNA. *Nature New Biology*. 236, 198–200 (1972).

173. Nucleotide sequence analysis of DNA: IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis. *Biochem. Biophys. Res. Commun.* 48, 1295–1302 (1972).

174. F. Sanger, J. E. Donelson, A. R. Coulson, H. Kössel, D. Fischer, Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc. Natl. Acad. Sci. U. S. A.* 70, 1209–1213 (1973).

175. R. Padmanabhan, E. Jay, R. Wu, Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage T4. *Proc. Natl. Acad. Sci. U. S. A.* 71, 2510–2514 (1974).

176. F. Sanger, A. R. Coulson, A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94, 441–448 (1975).

177. (available at https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.1975.0131).

178. F. Sanger, G. M. Air, B. G. Barrell, et al., Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*. 265, 687–695 (1977).

179. A. M. Maxam, W. Gilbert, A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*. 74, 560–564 (1977).

180. F. Sanger, S. Nicklen, A. R. Coulson, DNA sequencing with chain-terminating inhibitors. *Biochemistry*. 74, 104–108 (1977).

181. M. R. Atkinson, M. P. Deutscher, A. Kornberg, A. F. Russell, J. G. Moffatt, Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. *Biochemistry*. 8, 4897–4904 (1969).

182. 1982: GenBank Database Formed. *Genome.gov*, (available at

https://www.genome.gov/25520321/online-education-kit-1982-genbank-database-formed).

183. L. H. Augenlicht, D. Kobrin, Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer Res.* 42, 1088–1093 (1982).

184. K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, H. Erlich, Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb. Symp. Quant. Biol.* 51 Pt 1, 263–273 (1986).

185. L. M. Smith, J. Z. Sanders, R. J. Kaiser, et al., Fluorescence detection in automated DNA sequence analysis. *Nature*. 321, 674–679 (1986).

186. The Human Genome Project. *Genome.gov*, (available at https://www.genome.gov/human-genome-project).

187. 2001: First Draft of the Human Genome Sequence Released. *Genome.gov*, (available at https://www.genome.gov/25520483/online-education-kit-2001-first-draft-of-the-human-genome-sequence-released).

188. M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, P. Nyrén, Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242, 84–89 (1996).

189. M. Ronaghi, M. Uhlén, P. Nyrén, A sequencing method based on real-time pyrophosphate. *Science*. 281, 363, 365 (1998).

190. 454 Life Sciences obtains license from Pyrosequencing. *News Powered by Cision*, (available at https://news.cision.com/pyrosequencing/r/454-life-sciences-obtains-license-from-pyrosequencing,e83113).

191. Staff, 454 Life Sciences buys exclusive rights to genome sequencing technology - Boston Business Journal. *Boston Business Journal* (2003), (available at https://www.bizjournals.com/boston/blog/mass-high-tech/2003/08/454-life-sciences-buys-exclusive-rights.html).

192. M. Margulies, M. Egholm, W. E. Altman, et al., Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 437, 376–380 (2005).

193. S. Balasubramanian, Sequencing nucleic acids: from chemistry to medicine. *Chem. Commun.* . 47, 7281–7286 (2011).

194. S. Balasubramanian, Solexa sequencing: decoding genomes on a population scale. *Clin. Chem.* 61, 21–24 (2015).

195. C. Adessi, G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer, E. Kawashima, Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.* 28, E87 (2000).

196. (available at https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf).

197. M. L. Metzker, Sequencing technologies - the next generation. *Nat. Rev. Genet.* 11, 31–46 (2010).

198. M. I. Lefterova, C. J. Suarez, N. Banaei, B. A. Pinsky, Next-Generation Sequencing for Infectious Disease Diagnosis and Management. *The Journal of Molecular Diagnostics*. 17, 623–634 (2015).

199. brandonvd, About Us - Complete Genomics. *Complete Genomics*, (available at https://www.completegenomics.com/).

200. R. Drmanac, A. B. Sparks, M. J. Callow, et al., Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 327, 78–81 (2010).

201. MGI sequencing platforms: High-throughput gene sequencers, DNBSEQ™ sequencing technology-MGI, (available at https://en.mgitech.cn/products/).

202. D. Cyranoski, Direct Genomics revives Helicos sequencing system for China's hospitals. *Nat. Biotechnol.* 34, 122–123 (2016).

203. J. Eid, A. Fehr, J. Gray, et al., Real-time DNA sequencing from single polymerase molecules. *Science*. 323, 133–138 (2009).

204. M. J. Levene, J. Korlach, S. W. Turner, M. Foquet, H. G. Craighead, W. W. Webb, Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*. 299, 682–686 (2003).

205. Company history. *Oxford Nanopore Technologies*, (available at https://nanoporetech.com/about-us/history).

206. J. J. Kasianowicz, E. Brandin, D. Branton, D. W. Deamer, Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci. U. S. A.* 93, 13770–13773 (1996).

207. M. Akeson, D. Branton, G. Church, D. W. Deamer, Characterization of individual polymer molecules based on monomer-interface interactions. *US Patent* (2012), (available at https://patentimages.storage.googleapis.com/88/b9/af/8164402a4a4d0a/US20120160687A1.pdf).

208. S. Roy, W. A. LaFramboise, Y. E. Nikiforov, et al., Next-Generation Sequencing Informatics: Challenges and Strategies for Implementation in a Clinical Environment. *Archives of Pathology & Laboratory Medicine*. 140, 958–975, (2016).

209. Applied Biosystems Genetic Analysis Systems - GR (available at https://www.thermofisher.com/gr/en/home/life-science/sequencing/sanger-sequencing/sanger-sequencing-technology-accessories.html).

210. Sequencing Platforms | Compare NGS platform applications & specifications, (available at https://emea.illumina.com/systems/sequencing-platforms.html).

211. PacBio Sequel Systems - PacBio. *PacBio*, (available at https://www.pacb.com/products-and-services/sequel-system/).

212. Ion Torrent NGS Instruments - GR (available at https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-

torrent-next-generation-sequencing-run-sequence.html).

213. Product comparison. *Oxford Nanopore Technologies*, (available at https://nanoporetech.com/products/comparison).

214. B. Huang, A. Jennsion, D. Whiley, et al., Illumina sequencing of clinical samples for virus detection in a public health laboratory. *Scientific Reports*. 9 (2019), doi:10.1038/s41598-019-41830-w.

215. S. N. Naccache, J. Thézé, S. I. Sardi, et al., Distinct Zika Virus Lineage in Salvador, Bahia, Brazil. *Emerg. Infect. Dis.* 22, 1788–1792 (2016).

216. S. E. Eckert, J. Z.-M. Chan, D. Houniet, The Pathseek Consortium, J. Breuer, G. Speight, Enrichment by hybridisation of long DNA fragments for Nanopore sequencing. *Microb Genom*. 2, e000087 (2016).

217. C. C. S. Tan, S. Maurer-Stroh, Y. Wan, O. M. Sessions, P. F. de Sessions, A novel method for the capture-based purification of whole viral native RNA genomes. *AMB Express*. 9, 45 (2019).

218. J. Quick, N. D. Grubaugh, S. T. Pullan, et al., Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* 12, 1261–1276 (2017).

219. N. R. Faria, M. U. G. Kraemer, S. C. Hill, et al., Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*. 361, 894–899 (2018).

220. C. M. Hepp, J. H. Cocking, M. Valentine, et al., Phylogenetic analysis of West Nile Virus in Maricopa County, Arizona: Evidence for dynamic behavior of strains in two major lineages in the American Southwest. *PLOS ONE*. 13, e0205801 (2018).

221. L. Bragg, G. W. Tyson, Metagenomics Using Next-Generation Sequencing. *Methods in Molecular Biology*, 183–201 (2014).

222. R. D. Sleator, C. Shortall, C. Hill, Metagenomics. *Lett. Appl. Microbiol.* 47, 361–366 (2008).

223. C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, N. Segata, Corrigendum: Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* 35, 1211 (2017).

224. K. Bibby, Metagenomic identification of viral pathogens. *Trends in Biotechnology*. 31, 275–279 (2013).

225. M. J. Pallen, Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections. *Parasitology*. 141, 1856–1862 (2014).

226. G. M. Daly, N. Bexfield, J. Heaney, et al., A Viral Discovery Methodology for Clinical Biopsy Samples Utilising Massively Parallel Next Generation Sequencing. *PLoS ONE*. 6, p. e28879 (2011).

227. C. Kohl, A. Brinkmann, P. W. Dabrowski, A. Radonić, A. Nitsche, A. Kurth, Protocol for Metagenomic Virus Detection in Clinical Specimens1. *Emerging*

Infectious Diseases. 21 (2015).

228. A. Djikeng, R. Halpin, R. Kuzmickas, et al., Viral genome sequencing by random priming methods. *BMC Genomics*. 9, 5 (2008).

229. C. M. Malboeuf, X. Yang, P. Charlebois, et al., Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic Acids Research*. 41, e13–e13 (2013).

230. G. Rose, D. J. Wooldridge, C. Anscombe, E. T. Mee, R. V. Misra, S. Gharbia, Challenges of the Unknown: Clinical Application of Microbial Metagenomics. *International Journal of Genomics*. 2015, 1–10 (2015).

231. T. Rosseel, O. Ozhelvaci, G. Freimanis, S. Van Borm, Evaluation of convenient pretreatment protocols for RNA virus metagenomics in serum and tissue samples. *J. Virol. Methods*. 222, 72–80 (2015).

232. G. R. Reyes, J. P. Kim, Sequence-independent, single-primer amplification (SISPA) of complex DNA populations. *Molecular and Cellular Probes*. 5, 473–481 (1991).

233. G. R. Reyes, P. O. Yarbough, A. W. Tam, et al., Hepatitis E virus (HEV): the novel agent responsible for enterically transmitted non-A, non-B hepatitis. *Gastroenterol. Jpn.* 26 Suppl 3, 142–147 (1991).

234. S. M. Matsui, J. P. Kim, H. B. Greenberg, et al., The isolation and characterization of a Norwalk virus-specific cDNA. *J. Clin. Invest.* 87, 1456–1461 (1991).

235. P. R. Lambden, S. J. Cooke, E. O. Caul, I. N. Clarke, Cloning of noncultivatable human rotavirus by single primer amplification. *J. Virol.* 66, 1817–1822 (1992).

236. P. Froussard, rPCR: a powerful tool for random amplification of whole RNA sequences. *PCR Methods Appl.* 2, 185–190 (1993).

237. P. Froussard, A random-PCR method (rPCR) to construct whole cDNA library from low amounts of RNA. *Nucleic Acids Research*. 20, 2900–2900 (1992).

238. S. K. Bohlander, R. Espinosa, M. M. Le Beau, J. D. Rowley, M. O. Díaz, A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics*. 13, 1322–1324 (1992).

239. D. Wang, L. Coscoy, M. Zylberberg, P. C. Avila, H. A. Boushey, D. Ganem, J. L. DeRisi, Nonlinear partial differential equations and applications: Microarray-based detection and genotyping of viral pathogens. *Proceedings of the National Academy of Sciences*. 99, 15687–15692 (2002).

240. D. Wang, A. Urisman, Y.-T. Liu, M. Springer, et al., Viral Discovery and Sequence Recovery Using DNA Microarrays. *PLoS Biology*. 1, e2 (2003).

241. M. A. Marra, S. J. M. Jones, C. R. Astell, et al., The Genome sequence of the SARS-associated coronavirus. *Science*. 300, 1399–1404 (2003).

242. A. M. Gaynor, M. D. Nissen, D. M. Whiley, et al., Identification of a Novel Polyomavirus from Patients with Acute Respiratory Tract Infections. *PLoS*

Pathogens. 3, e64 (2007).

243. J. S. Towner, T. K. Sealy, M. L. Khristova, et al., Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog.* 4, e1000212 (2008).

244. D. L. Cox-Foster, S. Conlan, E. C. Holmes, et al., A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*. 318, 283–287 (2007).

245. T. Briese, J. T. Paweska, L. K. McMullan, et al., Genetic Detection and Characterization of Lujo Virus, a New Hemorrhagic Fever–Associated Arenavirus from Southern Africa. *PLoS Pathogens*. 5, e1000455 (2009).

246. G. Palacios, P.-L. Quan, O. J. Jabado, et al., Panmicrobial oligonucleotide array for diagnosis of infectious diseases. *Emerg. Infect. Dis.* 13, 73–81 (2007).

247. J. T. Paweska, N. H. Sewlall, T. G. Ksiazek, et al., Outbreak Control and Investigation Teams, Nosocomial outbreak of novel arenavirus infection, southern Africa. *Emerg. Infect. Dis.* 15, 1598–1602 (2009).

248. A. L. Greninger, E. C. Chen, T. Sittler, et al., A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS One*. 5, e13381 (2010).

249. J. G. Victoria, A. Kapoor, L. Li, et al., Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J. Virol.* 83, 4642–4651 (2009).

250. G. Grard, J. N. Fair, D. Lee, et al., A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. *PLoS Pathog.* 8, e1002924 (2012).

251. A. L. Greninger, S. N. Naccache, S. Federman, et al., Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis,Genome medicine 7.1, 99 (2015).

252. G. Watts, Lily Lyman Pinneo. *Lancet*. 380, 1552 (2012).

253. S. M. Buckley, J. Casals, W. G. Downs, Isolation and antigenic characterization of Lassa virus. *Nature*. 227, 174 (1970).

254. A. J. Dalton, W. P. Rowe, G. H. Smith, R. E. Wilsnack, W. E. Pugh, Morphological and cytochemical studies on lymphocytic choriomeningitis virus. *J. Virol.* 2, 1465–1478 (1968).

255. S. M. Buckley, J. Casals, Lassa Fever, a New Virus Disease of Man from West Africa. *Am. J. Trop. Med. Hyg.* 19, 680–691 (1970).

256. Nigeria Centre for Disease Control, (available at https://ncdc.gov.ng/diseases/factsheet/30).

257. Lassa Fever, WHO (available at https://www.who.int/docs/default-source/documents/emergencies/health-topics---lassa-fever/lassa-fever-introduction.pdf).

258. Lassa, VHFC (available at https://vhfc.org/diseases/lassa/).

259. J. B. McCormick, S. P. Fisher-Hoch, in *Arenaviruses I: The Epidemiology, Molecular and Cell Biology of Arenaviruses*, M. B. A. Oldstone, Ed. (Springer Berlin Heidelberg, Berlin, Heidelberg), 75–109 (2002).

260. Lassa Fever, WHO, Regional Office for Africa (available at https://www.afro.who.int/health-topics/lassa-fever/#tab=tab_2).

261. Lassa Fever Fact Sheet, WHO (available at https://www.who.int/news-room/fact-sheets/detail/lassa-fever).

262. CDC WHO, Technical guidelines for integrated disease surveillance and response in the African region. *Brazzaville, Republic of Congo and Atlanta, USA*, 1–398 (2010).

263. Lassa Fever, WHO, Regional Office for Africa (available at https://www.afro.who.int/health-topics/lassa-fever).

264. Viral hemorrhagic fevers (available at https://nuh.nhs.uk/viral-hemorrhagic-fever/).

265. CDC (available at https://wwwnc.cdc.gov/travel/yellowbook/2020/travel-related-infectious-diseases/viral-hemorrhagic-fevers).

266. S. Olschläger, S. Günther, Rapid and specific detection of Lassa virus by reverse transcription-PCR coupled with oligonucleotide array hybridization. *J. Clin. Microbiol.* 50, 2496–2499 (2012).

267. S. Olschlager, M. Lelke, P. Emmerich, et al., Improved Detection of Lassa Virus by Reverse Transcription-PCR Targeting the 5' Region of S RNA. *Journal of Clinical Microbiology*. 48, 2009–2013 (2010).

268. K. Lunkenheimer, F. T. Hufert, H. Schmitz, Detection of Lassa virus RNA in specimens from patients with Lassa fever by using the polymerase chain reaction. *J. Clin. Microbiol.* 28, 2689–2692 (1990).

269. A. H. Demby, J. Chamberlain, D. W. Brown, C. S. Clegg, Early diagnosis of Lassa fever by reverse transcription-PCR. *J. Clin. Microbiol.* 32, 2898–2903 (1994).

270. RealStar® Lassa Virus RT-PCR Kit - Altona-Diagnostics EN, (available at https://altona-diagnostics.com/en/products/reagents-140/reagents/realstar-real-time-pcr-reagents/realstar-lassavirus-rt-pcr-kit-ce.html).

271. J. Kingdon, M. J. Largen, The kingdom field guide to African mammals. *Zool. J. Linn. Soc.* 120, 479 (1997).

272. C. G. Coetzee, The biology, behaviour, and ecology of Mastomys natalensis in southern Africa. *Bull. World Health Organ.* 52, 637–644 (1975).

273. R. H. Makundi, A. W. Massawe, L. S. Mulungu, Reproduction and population dynamics of Mastomys natalensis Smith, 1834 in an agricultural landscape in the Western Usambara Mountains, Tanzania. *Integr. Zool.* 2, 233–238 (2007).

274. L. S. Karan, M. T. Makenov, M. G. Korneev, et al., Lassa Virus in the Host Rodent Mastomys Natalensis within Urban Areas of N'zerekore, Guinea,

BioRxiv, 616466 (2019)

275. D. S. Grant, H. Khan, J. Schieffelin, D. G. Bausch, in *Emerging Infectious Diseases*, Ö. Ergönül, F. Can, L. Madoff, M. Akova, Eds. (Academic Press, Amsterdam), 37–59 (2014).

276. T. P. Monath, Lassa fever: review of epidemiology and epizootiology. *Bull. World Health Organ.* 52, 577–592 (1975).

277. R. A. Keenlyside, J. B. McCormick, P. A. Webb, E. Smith, L. Elliott, K. M. Johnson, Case-control study of Mastomys natalensis and humans in Lassa virus-infected households in Sierra Leone. *Am. J. Trop. Med. Hyg.* 32, 829–837 (1983).

278. J. B. McCormick, P. A. Webb, J. W. Krebs, K. M. Johnson, E. S. Smith, A prospective study of the epidemiology and ecology of Lassa fever. *J. Infect. Dis.* 155, 437–444 (1987).

279. T. P. Monath, V. F. Newhouse, G. E. Kemp, H. W. Setzer, A. Cacciapuoti, Lassa virus isolation from Mastomys natalensis rodents during an epidemic in Sierra Leone. *Science*. 185, 263–265 (1974).

280. A. H. Demby, A. Inapogui, K. Kargbo, et al., Lassa fever in Guinea: II. Distribution and prevalence of Lassa virus infection in small mammals. *Vector Borne Zoonotic Dis.* 1, 283–297 (2001).

281. A. Olayemi, D. Cadar, N. 'faly Magassouba, et al., New Hosts of The Lassa Virus. *Sci. Rep.* 6, 25280 (2016).

282. Lassa fever. *World Health Organization*, (available at http://www.who.int/news-room/fact-sheets/detail/lassa-fever).

283. M. O. Iroezindu, U. S. Unigwe, C. C. Okwara, et al., Lessons learnt from the management of a case of Lassa fever and follow-up of nosocomial primary contacts in Nigeria during Ebola virus disease outbreak in West Africa. *Tropical Medicine & International Health*. 20, 1424–1430 (2015).

284. S. P. Fisher-Hoch, O. Tomori, A. Nasidi, G. I. Perez-Oronoz, Y. Fakile, L. Hutwagner, J. B. McCormick, Review of cases of nosocomial Lassa fever in Nigeria: the high price of poor medical practice. *BMJ*. 311, 857–859 (1995).

285. D. G. Bausch, P. E. Rollin, A. H. Demby, et al., Diagnosis and clinical virology of Lassa fever as evaluated by enzyme-linked immunosorbent assay, indirect fluorescent-antibody test, and virus isolation. *J. Clin. Microbiol.* 38, 2670–2677 (2000).

286. E. Fichet-Calvet, S. Ölschläger, T. Strecker et al., Spatial and temporal evolution of Lassa virus in the natural host population in Upper Guinea. *Sci. Rep.* 6, 21977 (2016).

287. J. Mariën, B. Borremans, F. Kourouma, et al., Evaluation of rodent control to fight Lassa fever based on field data and mathematical modelling. *Emerg. Microbes Infect.* 8, 640–649 (2019).

288. J. D. Frame, D. J. Gocke, J. M. Baldwin, J. M. Troup, Lassa Fever, a New Virus

Disease of Man from West Africa. *The American Journal of Tropical Medicine and Hygiene*. 19, 670–676 (1970).

289. D. E. Carey, G. E. Kemp, H. A. White, et al., Lassa fever. Epidemiological aspects of the 1970 epidemic, Jos, Nigeria. *Trans. R. Soc. Trop. Med. Hyg.* 66, 402–408 (1972).

290. J. D. Frame, P. B. Jahrling, J. E. Yalley-Ogunro, M. H. Monson, Endemic Lassa fever in Liberia. II. Serological and virological findings in hospital patients. *Trans. R. Soc. Trop. Med. Hyg.* 78, 656–660 (1984).

291. J. Knobloch, J. B. McCormick, P. A. Webb, M. Dietrich, H. H. Schumacher, E. Dennis, Clinical observations in 42 patients with Lassa fever. *Tropenmed. Parasitol.* 31, 389–398 (1980).

292. T. P. Monath, M. Maher, J. Casals, R. E. Kissling, A. Cacciapuoti, Lassa Fever in the Eastern Province of Sierra Leone, 1970–1972. *The American Journal of Tropical Medicine and Hygiene*. 23, 1140–1149 (1974).

293. S. L. M. Whitmer, T. Strecker, D. Cadar, et al., New Lineage of Lassa Virus, Togo, 2016. *Emerg. Infect. Dis.* 24, 599–602 (2018).

294. D. Safronetz, J. E. Lopez, N. Sogoba, et al., Detection of Lassa virus, Mali. *Emerg. Infect. Dis.* 16, 1123–1126 (2010).

295. L. Kouadio, K. Nowak, C. Akoua-Koffi, et al., Lassa Virus in Multimammate Rats, Côte d'Ivoire, 2013. *Emerg. Infect. Dis.* 21, 1481–1483 (2015).

296. D. U. Ehichioya, S. Dellicour, M. Pahlmann, et al., Phylogeography of Lassa Virus in Nigeria. *J. Virol.* 93 (2019)

297. C.-M. Swaan, P.-J. van den Broek, S. Wijnands, J. E. van Steenbergen, Management of viral haemorrhagic fever in the Netherlands. *Euro Surveill.* 7, 48–50 (2002).

298. D. U. Ehichioya, M. Hass, B. Becker-Ziaja, et al., Current molecular epidemiology of Lassa virus in Nigeria. *J. Clin. Microbiol.* 49, 1157–1161 (2011).

299. E. K. Dzotsi, S.-A. Ohene, F. Asiedu-Bekoe, et al., The first cases of Lassa fever in Ghana. *Ghana Med. J.* 46, 166–170 (2012).

300. M. Mateo, C. Picard, Y. Sylla, et al., Fatal Case of Lassa Fever, Bangolo District, Côte d'Ivoire, 2015. *Emerging Infectious Diseases*. 25, 1753–1756 (2019).

301. N. Sogoba, H. Feldmann, D. Safronetz, Lassa fever in West Africa: evidence for an expanded region of endemicity. *Zoonoses Public Health*. 59 Suppl 2, 43–47 (2012).

302. J. T. Manning, N. Forrester, S. Paessler, Lassa virus isolates from Mali and the Ivory Coast represent an emerging fifth lineage. *Front. Microbiol.* 6, 1037 (2015).

303. L. E. Kafetzopoulou, S. T. Pullan, P. Lemey, et al., Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science*. 363, 74–77 (2019).

304. K. Lewandowski, A. Bell, R. Miles, et al., The Effect of Nucleic Acid Extraction

Platforms and Sample Storage on the Integrity of Viral RNA for Use in Whole Genome Sequencing. *J. Mol. Diagn.* 19, 303–312 (2017).

305. C. J. Edwards, S. R. Welch, J. Chamberlain, R. Hewson, H. Tolley, P. A. Cane, G. Lloyd, Molecular diagnosis and analysis of Chikungunya virus. *J. Clin. Virol.* 39, 271–275 (2007).

306. C. Drosten, S. Göttig, S. Schilling, M. Asper, M. Panning, H. Schmitz, S. Günther, Rapid detection and quantification of RNA of Ebola and Marburg viruses, Lassa virus, Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus, dengue virus, and yellow fever virus by real-time reverse transcription-PCR. *J. Clin. Microbiol.* 40, 2323–2330 (2002).

307. S. Nikisins, T. Rieger, P. Patel, R. Müller, S. Günther, M. Niedrig, International external quality assessment study for molecular detection of Lassa virus. *PLoS Negl. Trop. Dis.* 9, e0003793 (2015).

308. A. L. Greninger, S. N. Naccache, S. Federman, et al., Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome Med.* 7, 99 (2015).

309. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25, 1754–1760 (2009).

310. Burrows-Wheeler Aligner, *GitHub* (available at http://github.com/lh3/bwa).

311. H. Li, Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. *Bioinformatics*. 28, 1838–1844 (2012).

312. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013), (available at http://arxiv.org/abs/1303.3997).

313. Burrows-Wheeler Aligner, (available at http://bio-bwa.sourceforge.net/).

314. Burrows-Wheeler Aligner Manual page, (available at http://bio-bwa.sourceforge.net/bwa.shtml).

315. SAMtools, (available at http://samtools.sourceforge.net/).

316. SAMtools Manual page, (available at http://www.htslib.org/doc/samtools-1.2.html).

317. bedtools: a powerful toolset for genome arithmetic — bedtools 2.27.0 documentation, (available at https://bedtools.readthedocs.io/en/latest/).

318. I. Milne, G. Stephen, M. Bayer, P. J. A. Cock, L. Pritchard, L. Cardle, P. D. Shaw, D. Marshall, Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* 14, 193–202 (2013).

319. A. R. Penedos, R. Myers, B. Hadef, F. Aladin, K. E. Brown, Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks. *PLoS One*. 10, e0143081 (2015).

320. N. J. Loman, J. Quick, J. T. Simpson, A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*. 12, 733–735 (2015).

321. zibraproject, zika-pipeline, GitHub (available at https://github.com/zibraproject/zika-pipeline).

322. J. C. Dohm, C. Lottaz, T. Borodina, H. Himmelbauer, SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.* 17, 1697–1706 (2007).

323. A. Bankevich, S. Nurk, D. Antipov, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19, 455–477 (2012).

324. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation (2016).

325. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* 15, R46 (2014).

326. D. Kim, L. Song, F. P. Breitwieser, S. L. Salzberg, Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.* 26, 1721–1729 (2016).

327. Porechop. *GitHub*, (available at https://github.com/rrwick/Porechop).

328. seqtk. *GitHub*, (available at https://github.com/lh3/seqtk).

329. The GNU Awk User's Guide, (available at https://www.gnu.org/software/gawk/manual/gawk.html).

330. bioawk. *GitHub*, (available at https://github.com/lh3/bioawk).

331. FASTQ-SAMPLE, (available at https://homes.cs.washington.edu/~dcjones/fastq-tools/fastq-sample.html).

332. N. J. Loman, A. R. Quinlan, Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics*. 30, 3399–3401 (2014).

333. M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*. 27, 578–579 (2011).

334. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736 (2017).

335. C. M. Malboeuf, X. Yang, P. Charlebois, et al., Complete viral RNA genome sequencing of ultra-low copy samples by sequence-independent amplification. *Nucleic Acids Res.* 41, e13 (2013).

336. A. Bell, K. Lewandowski, R. Myers, et al., Genome sequence analysis of Ebola virus in clinical samples from three British healthcare workers, August 2014 to March 2015. *Eurosurveillance*. 20, 21131 (2015).

337. FASTQ-SAMPLE, GitHub (available at https://github.com/fplaza/fastq-sample).

338. S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, A. M. Phillippy, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736 (2017).

339. L. E. Kafetzopoulou, K. Efthymiadis, K. Lewandowski, et al., Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. *Euro Surveill.* 23 (2018).

340. WHO | Lassa Fever – Nigeria (2018) (available at http://www.who.int/csr/don/23-march-2018-lassa-fever-nigeria/en/).

341. Nigeria Centre for Disease Control, (available at https://ncdc.gov.ng/news/121/early-results-of-lassa-virus-sequencing-%26-implications-for-current-outbreak-response-in-nigeria).

342. Philippe Lemey, 2018 LASV sequencing. *Virological* (2018), (available at http://virological.org/t/2018-lasv-sequencing/180).

343. Philippe Lemey, 2018 LASV sequencing, continued. *Virological* (2018), (available at http://virological.org/t/2018-lasv-sequencing-continued/192).

344. M. C. Walter, K. Zwirglmaier, P. Vette, S. A. Holowachuk, K. Stoecker, G. H. Genzel, M. H. Antwerpen, MinION as part of a biomedical rapidly deployable laboratory. *J. Biotechnol.* 250, 16–22 (2017).

345. B. Daniel, D. D. W, *Nanopore Sequencing: An Introduction* (World Scientific, 2019).

Publications

Research article

Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples

Liana E. Kafetzopoulou[1,2], Kyriakos Efthymiadis[3], Kuiama Lewandowski[1], Ant Crook[1], Dan Carter[1,2], Jane Osborne[4], Emma Aarons[4], Roger Hewson[1,2], Julian A. Hiscox[2,5], Miles W. Carroll[1,2], Richard Vipond[1,2], Steven T. Pullan[1,2]

1. Public Health England, National Infections Service, Porton Down, United Kingdom
2. NIHR Health Protection Research Unit in Emerging and Zoonotic Infections, Liverpool, United Kingdom
3. Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium
4. Rare and Imported Pathogens Laboratory, Public Health England, Porton Down, United Kingdom
5. Institute of Infection and Global Health, University of Liverpool, United Kingdom

Correspondence: Steven Pullan (steven.pullan@phe.gov.uk)

Citation style for this article:
Kafetzopoulou Liana E., Efthymiadis Kyriakos, Lewandowski Kuiama, Crook Ant, Carter Dan, Osborne Jane, Aarons Emma, Hewson Roger, Hiscox Julian A., Carroll Miles W., Vipond Richard, Pullan Steven T.. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. Euro Surveill. 2018;23(50):pii=1800228. https://doi.org/10.2807/1560-7917.ES.2018.23.50.1800228

Article submitted on 27 Apr 2018 / accepted on 23 Oct 2018 / published on 13 Dec 2018

Background: The recent global emergence and re-emergence of arboviruses has caused significant human disease. Common vectors, symptoms and geographical distribution make differential diagnosis both important and challenging. **Aim:** To investigate the feasibility of metagenomic sequencing for recovering whole genome sequences of chikungunya and dengue viruses from clinical samples. **Methods:** We performed metagenomic sequencing using both the Illumina MiSeq and the portable Oxford Nanopore MinION on clinical samples which were real-time reverse transcription-PCR (qRT-PCR) positive for chikungunya (CHIKV) or dengue virus (DENV), two of the most important arboviruses. A total of 26 samples with a range of representative clinical Ct values were included in the study. **Results:** Direct metagenomic sequencing of nucleic acid extracts from serum or plasma without viral enrichment allowed for virus identification, subtype determination and elucidated complete or near-complete genomes adequate for phylogenetic analysis. One PCR-positive CHIKV sample was also found to be coinfected with DENV. **Conclusions:** This work demonstrates that metagenomic whole genome sequencing is feasible for the majority of CHIKV and DENV PCR-positive patient serum or plasma samples. Additionally, it explores the use of Nanopore metagenomic sequencing for DENV and CHIKV, which can likely be applied to other RNA viruses, highlighting the applicability of this approach to front-line public health and potential portable applications using the MinION.

Introduction

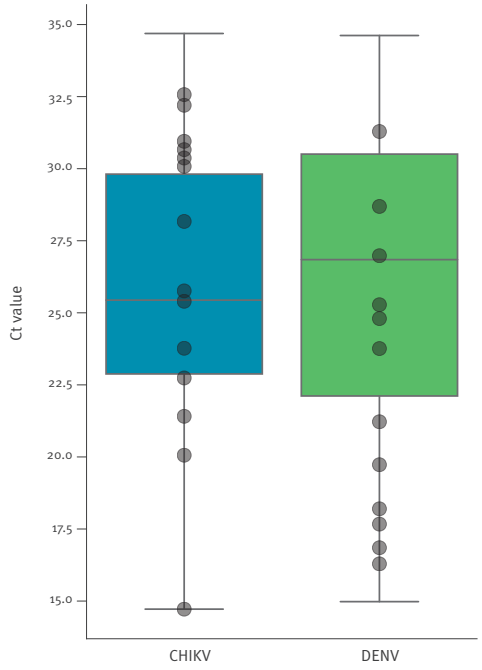
Arboviruses are predominantly RNA viruses that replicate in haematophagous (blood-sucking) arthropod vectors such as ticks, mosquitoes and other biting flies to maintain their transmission cycle [1]. Human disease outbreaks caused by arboviruses have increased in prevalence since the 2000's, led by the spread of mosquito-borne arboviruses such as chikungunya (CHIKV), dengue (DENV), West Nile (WNV), yellow fever (YFV) and Zika (ZIKV) viruses across both hemispheres [2]. CHIKV and DENV are of particular global health concern, as they have lost the need for enzootic amplification and consequently have caused extensive epidemics [3].

CHIKV is a single-stranded positive-sense RNA virus of the alphavirus genus, which causes the debilitating arthritic disease, chikungunya [4]. It has spread globally and been designated a serious emerging disease by the World Health Organization [5]. Outbreaks of CHIKV since 2005 have been associated with increased morbidity and possibly mortality [6,7].

DENV, which causes dengue, is a single-stranded positive-sense RNA virus of the flavivirus genus and the most prevalent human arboviral pathogen. Dengue occurs following infection with one of four DENV serotypes (DENV1–4). A minority of cases develop acute haemorrhagic manifestations and multi-organ failure. Despite DENV cases being under-reported, a 143.1% increased global incidence was estimated between 2005 and 2015 [8]. Approximately 500,000 DENV infected patients worldwide require hospitalisation annually [9].

Figure 1

Cycle threshold (Ct) values distribution of chikungunya (n = 73) and dengue virus (n = 368) positive samples from the Rare and Imported Pathogens Laboratory, Public Health England, United Kingdom, 2016 (n = 441 total samples)



CHIKV: chikungunya virus; Ct: cycle threshold; DENV: dengue virus.

The 14 CHIKV and 12 DENV samples selected for this work are indicated by circles. For each virus, the median Ct value of positive samples by quantitative real-time PCR is shown (horizontal line inside the box), as well as 25th and 75th percentiles (box lower and upper boundaries) and total range (whiskers).

Both CHIKV and DENV are predominantly transmitted to humans via *Aedes* species mosquitoes, particularly *Ae. aegypti* and *Ae. albopictus* [10,11], and share clinical presentations of arthralgia, headache, high fever, myalgia and rash. Circulation of CHIKV, DENV (and other arboviruses) in the same areas leads to challenges in differential diagnosis, especially in endemic regions in which diagnosis is predominantly symptom-based [12]. Additionally, reports of arboviral coinfections are increasingly common [13-16].

Metagenomic RNA sequencing allows for identification of multiple pathogens within a sample in a non-targeted and unbiased manner. It has identified causative agents in outbreaks, e.g. Lujo virus in South Africa [17], Bundibugyo ebolavirus in Uganda [18] and lead to novel virus discovery such as a rhabdovirus causing haemorrhagic fever in central Africa [19]. It also provides genomic information for typing and surveillance. Real-time genomic surveillance was facilitated on-site by the portable Oxford Nanopore MinION sequencer during the 2014–16 Ebola virus (EBOV) epidemic in West Africa and the 2015-16 ZIKV outbreak in the Americas [20-23] for epidemiological and transmission chain investigations [24]. In both examples, an amplicon sequencing approach was used, but viruses and bacteria from clinical, environmental and vector samples have been sequenced using metagenomic approaches on the MinION [25-28]. Metagenomic sequencing of CHIKV was demonstrated in principle on the MinION by Greninger et al. in 2015 reporting the detection of CHIKV from a human blood sample [28]. Additionally, Illumina-based metagenomics identified CHIKV coinfections within a ZIKV sample cohort [29], with the high proportion of CHIKV reads present making it a promising target for the approach.

In this study we set out to test the feasibility of direct metagenomic sequencing of DENV and CHIKV genomes from a cohort of clinical serum and plasma samples across a representative range of viral loads. The objective was to assess the proportion of viral nucleic acid relative to patient/background present in each sample and determine the sequencing limits for whole genome retrieval using both the laboratory-based Illumina technology and the portable MinION platform.

Methods

Sample collection and nucleic acid extraction

Twenty-six routine diagnostic samples, nine plasma and 17 serum, were obtained from the Rare and Imported Pathogens Laboratory (RIPL), Public Health England (PHE), Porton Down. All had previously tested positive by real-time reverse transcription-PCR (qRT-PCR) for chikungunya or dengue virus, with a maximum cut-off value of cycle threshold (Ct) 35. These samples had been selected based on their Ct values, among a larger set of 441 samples, so as to represent a Ct clinical range. Total nucleic acid was extracted from 140 µL of each using the QIAamp viral RNA kit (Qiagen, Hilden, Germany) replacing carrier RNA with linear polyacrylamide and eluting in 60 µL elution buffer provided in the kit, followed by treatment with TURBO DNase (Thermo Fisher Scientific, Waltham, United States (US)) at 37 °C for 30 min. RNA was purified and concentrated to 8 µL using the RNA Clean and Concentrator-5 kit (Zymo Research, Irvine, US).

Molecular confirmation and quantification

Drosten et al. [30] and Edwards et al. [31] RT-PCR assays were used for confirmation of DENV and CHIKV respectively. RNA oligomers were used as standards for genome copy quantitation.

Table 1

Description of samples positive for chikungunya and dengue virus by real-time reverse transcription-PCR with corresponding Illumina mapping data, United Kingdom, 2017[a] (n = 26 samples)

Sample	Ct value	Estimated genome copy number in the sample (/mL)	Sample type	Total reads (R1 + R2)[b]	% reads mapping to reference viral genome	% 20x coverage	% 10x coverage	Reference virus[c]	Reference size (nts)
CHIKV 1	14.72	2.12E+10	Plasma	1,113,560	78.32	99.59	99.72	CHIKV	11,826
CHIKV 2	20.06	5.49E+08	Serum	1,278,624	98.48	99.14	99.47	CHIKV	11,826
CHIKV 3	21.41	2.18E+08	Plasma	1,391,258	95.23	98.86	99.37	CHIKV	11,826
CHIKV 4	22.74	8.76E+07	Plasma	888,968	19.16	97.08	97.32	CHIKV	11,826
CHIKV 5	23.77	4.33E+07	Plasma	1,357,606	97.13	99.16	99.58	CHIKV	11,826
CHIKV 6	25.4	1.42E+07	Serum	3,236,848	34.88	97.80	98.40	CHIKV	11,826
CHIKV 7	25.76	1.11E+07	Plasma	3,748,070	72.77	99.04	99.56	CHIKV	11,826
CHIKV 8	28.17	2.13E+06	Plasma	1,499,952	28.41	98.69	99.00	CHIKV	11,826
CHIKV 9	30.08	5.76E+05	Serum	1,035,026	6.66	95.98	98.22	CHIKV	11,826
CHIKV 10	30.37	4.72E+05	Serum	1,575,222	16.84	97.39	98.01	CHIKV	11,826
CHIKV 11	30.66	3.87E+05	Serum	1,143,054	13.52	95.36	96.96	CHIKV	11,826
CHIKV 12	30.95	3.17E+05	Serum	1,507,380	10.93	96.11	96.52	CHIKV	11,826
CHIKV 13	32.2	1.35E+05	Serum	1,323,920	5.03	88.47	89.38	CHIKV	11,826
CHIKV 14	32.57	1.05E+05	Serum	1,479,404	21.72	96.32	96.93	CHIKV	11,826
DENV 1	16.29	4.21E+09	Plasma	439,292	93.44	99.51	99.58	DENV 1	10,735
DENV 2	16.85	2.83E+09	Serum	513,472	92.56	99.40	99.58	DENV 1	10,735
DENV 3	17.67	1.58E+09	Plasma	738,814	92.53	99.58	99.58	DENV 2	10,723
DENV 4	18.20	1.09E+09	Serum	477,368	93.97	98.73	99.12	DENV 2	10,723
DENV 5	19.73	3.67E+08	Serum	915554	89.65	99.14	99.40	DENV 2	10,723
DENV 6	21.22	3.61E+07	Serum	3,587,926	83.87	99.68	99.69	DENV 4	10,649
DENV 7	23.76	2.11E+07	Serum	4,146,678	2.17	86.99	89.13	DENV 1	10,735
DENV 8	24.8	1.01E+07	Serum	777,264	69.23	99.56	99.58	DENV 3	10,707
DENV 9	25.28	7.17E+06	Plasma	787,728	26.97	98.77	98.81	DENV 2	10,723
DENV 10	26.98	2.15E+06	Serum	596,240	6.58	93.47	93.97	DENV 3	10,707
DENV 11	28.69	6.39E+05	Serum	1,034,698	3.73	94.44	94.70	DENV 1	10,735
DENV 12	31.29	1.01E+05	Serum	1,374,766	0.47	71.46	77.76	DENV 1	10,735

CHIKV: chikungunya virus; Ct: cycle threshold; DENV: dengue virus.

[a] The Illumina mapping data presented in the table were obtained in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

[b] 'R1 + R2' indicates paired-end sequencing.

[c] For DENV the serotype is also indicated.

Metagenomic cDNA reparation

Complementary DNA (cDNA) was prepared using a Sequence Independent Single Primer Amplification (SISPA) approach adapted from Greninger et al. [28]. Reverse transcription and second strand cDNA synthesis were as described [28]. cDNA amplification was performed using AccuTaq LA (Sigma, Poole, United Kingdom), in which 5µL of cDNA and 1 µL (100 pmol/µL) Primer B (5'-GTTTCCCACTGGAGGATA-3') were added to a 50 µL reaction, according to manufacturer's instructions. PCR conditions were 98 °C for 30s, followed by 30 cycles of 94 °C for 15 s, 50 °C for 20 s, and 68 °C for 5 min, and a final step of 68 °C for 10 min. Amplified cDNA was purified using a 1:1 ratio of AMPure XP beads (Beckman Coulter, Brea, California (CA)) and quantified using the Qubit High Sensitivity dsDNA kit (Thermo Fisher, Waltham, US).

MinION library preparation and sequencing

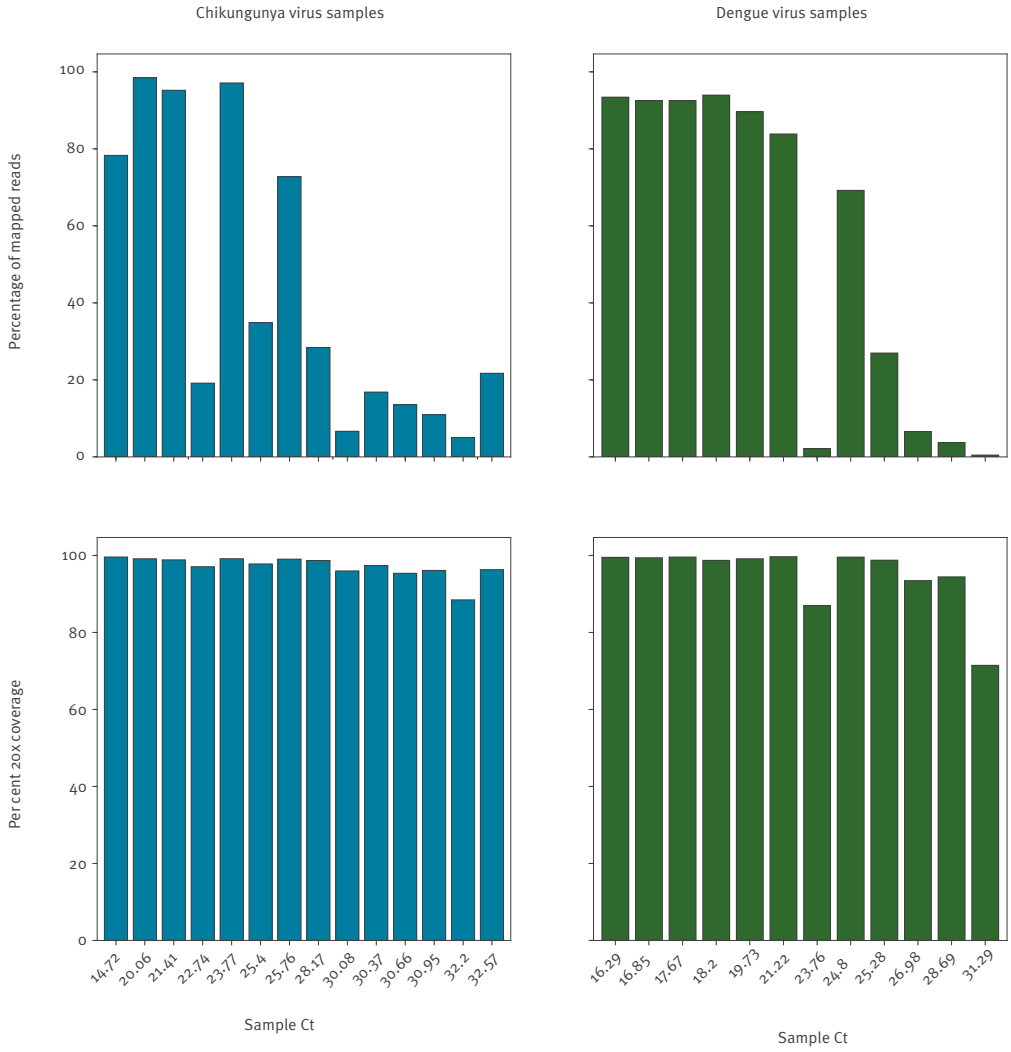
MinION sequencing libraries were prepared using total amplified cDNA of each sample to a maximum of 1 µg. Oxford Nanopore kits SQK-NSK007 or SQK-LSK208 (2D), SQK-LSK308 (1D²) and SQK-RBK001 (Rapid) were used and each sample was run individually on the appropriate flow cell (FLO-MIN105, FLO-MIN106 or FLO-MIN107) using the 48hr run script. Base calling was performed using Metrichor (ONT) for SQK-NSK007 and SQK-LSK208 or Albacore v1.2 for SQK-LSK308 and SQK-RBK001. Poretools [32] was used to extract FASTQ files from Metrichor FAST5 files.

Illumina library preparation and sequencing

Nextera XT V2 kit (Illumina) sequencing libraries were prepared using 1.5 ng of amplified cDNA as per manufacturer's instructions. Samples were multiplexed in batches of a maximum of 16 samples per run and

Figure 2

Proportion of reads mapping to the appropriate viral reference sequence and proportion of reference genome sequenced at minimum 20-fold coverage in each chikungunya or dengue virus positive sample, United Kingdom, 2017[a] (n = 26 samples)



Ct: cycle threshold.

[a] The Illumina data presented in the figure were obtained in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

The percentage of total reads mapping to the appropriate reference sequence is plotted in the upper panel. Lower panels display the percentage of the reference genome sequenced to a minimum depth of 20-fold in the Illumina data.

Table 2

Description of chikungunya and dengue virus positive samples by real-time reverse transcription-PCR and corresponding Nanopore sequencing data, United Kingdom, 2017[a] (n = 8 samples)

Sample	Ct value	cDNA amount used for the library (ng)	Sequencing kit (2D kit version)	Flow cell (FLO-)	1D total bp	1D total reads	1D mean read length (nt)	1D max read length (nt)
CHIKV 1	14.7	431.5	SQK-NSK007	MIN105	1.51E+08	267,171	564	92,712
CHIKV 3	21.4	928.8	SQK-LSK208	MIN106	1.63E+09	1,891,028	862	99,031
CHIKV 4	22.7	113.4	SQK-NSK007	MIN105	1.74E+08	216,493	805	125,387
CHIKV 9	30.1	212.4	SQK-LSK208	MIN106	2.12E+09	3,481,358	608	121,711
DENV 1	16.3	1,626.0	SQK-NSK007	MIN105	2.42E+08	284,622	851	115,494
DENV 2	16.9	1,626.0	SQK-NSK007	MIN105	1.55E+08	203,700	760	52,157
DENV 6	21.2	475.0	SQK-LSK208	MIN106	1.22E+09	1,377,721	886	118,733
DENV 11	28.7	65.8	SQK-LSK208	MIN106	7.07E+08	1,111,566	636	119,438

Ct: cycle threshold.

[a] The Nanopore data presented in the table were obtained in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

sequenced on a 2x150 bp-paired end Illumina MiSeq run, by Genomics Services Development Unit, PHE.

Data handling

BWA MEM v0.7.15 [33] was used to align reads to the following references (GenBank ID): DENV Serotype 1 (NC_001477.1), DENV Serotype 2 (NC_001474.2), DENV Serotype 3 (NC_001475.2), DENV Serotype 4 (NC_002640.1) and CHIKV (NC_004162.2) using -x ont2d mode for Nanopore and MEM defaults for Illumina reads. Samtools v1.4 [34] was used to compute percentage reads mapped and coverage depth. Bedtools v2.26.0 [35] was used to calculate genome coverage at 10x and 20x. Mapping consensuses for Illumina were generated using in-house software QuasiBam [36] and for MinION using a simple pileup with bases called at a minimum depth of 20x and 70% support fraction. Nanopolish variants [24,37] was used in consensus mode to compute an error-corrected consensus sequence from the Rapid kit data. Taxonomic classification was performed using Kraken (0.10.4-beta) [38] and a locally built database populated with all RefSeq bacterial, viral, and archaeal genomes plus additional sequences [39]. De novo assemblies were generated using Spades 3.8.2 [40] in combination with SSPACE Standard v3.0 [41] for Illumina generated sequences and Canu v1.6 [41,42] for Nanopore sequences (settings: corOutCoverage = 1,000; genomeSize = 12,000; minReadLength = 300, minOverlapLength = 50).

Consensus sequences for all samples tested are available in Genbank, raw fast5 files from 1D2 and 1D data (viral reads only) are deposited in SRA (Both under BioProject PRJNA508296).

Results

Metagenomic Illumina sequencing

A total of 73 samples tested during 2016 in RIPL diagnostic laboratories, PHE Porton Down, were positive by qRT-PCR for CHIKV, and 368 were positive for DENV. Median Ct for CHIKV was 26.1, for DENV it was 26.8. For each virus, samples representing the range of viral titres seen during 2016 were selected, based on qRT-PCR Ct value (Figure 1). CHIKV samples selected (n = 14) ranged from Ct 14.72 to Ct 32.57, corresponding to 10^{10} and 10^{5} genome copies per mL of plasma or serum. DENV samples selected (n = 12) ranged from Ct 16.29 to Ct 31.29, corresponding to 10^{9} and 10^{5} genome copies per mL (Table 1). To measure the proportion of viral nucleic acid present relative to host/background and assess genome coverage, all samples were processed as described in methods and Illumina sequenced (Table 1). The proportion of total reads mapping to the respective viral reference was high for both viruses (Figure 2). In some low Ct samples, over 90% of reads mapped to the viral reference and proportions over 50% were still observed at mid-Ct range. The lowest proportions observed were 5.03% and 0.47% for CHIKV and DENV respectively (Table 1, Figure 2). The majority of samples returned over 95% genome coverage at 20x (21/26 samples) and over 98% genome coverage at 10x (20/26 samples). Irrespective of lower mapping percentages in high Ct value samples, genome coverage of 88.5% (20x) and 89.4% (10x) for CHIKV and 75.0% (20x) and 77.8% (10x) for DENV was observed.

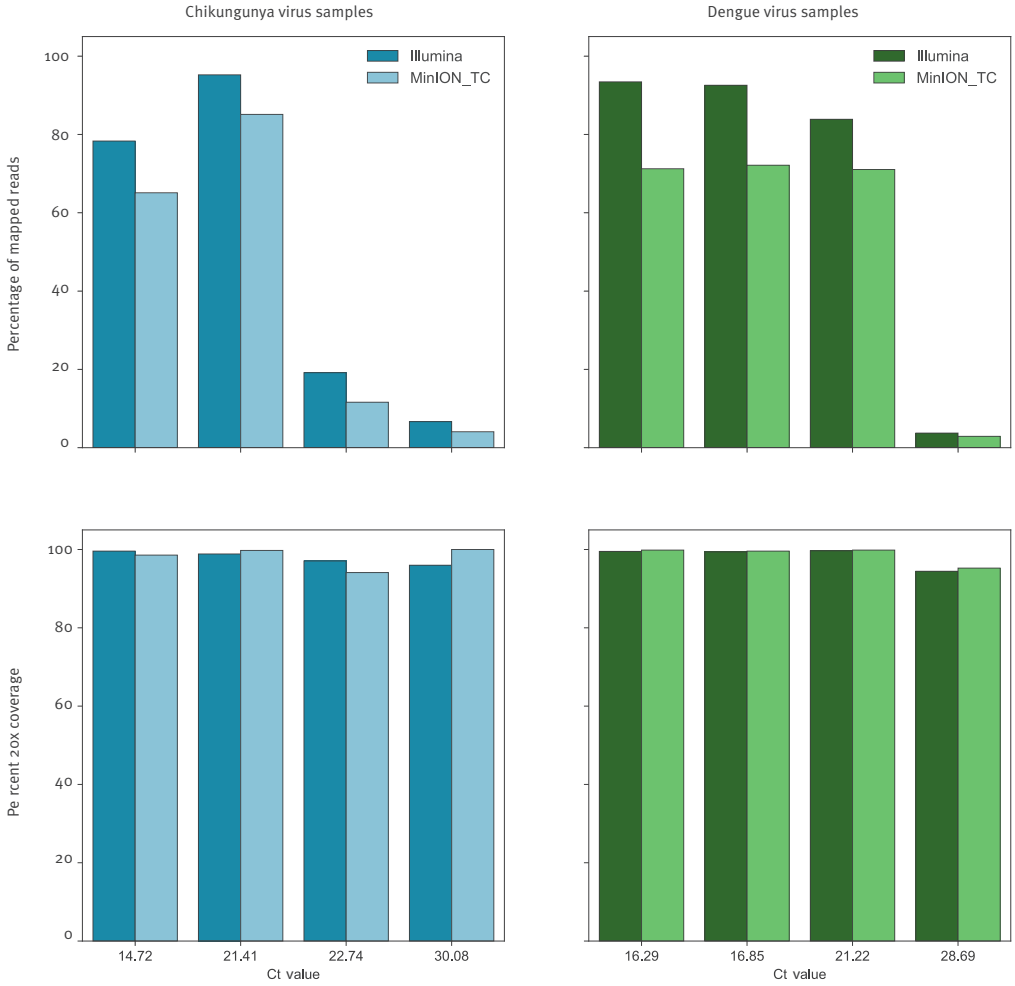
Metagenomic MinION sequencing

Four representative samples for each virus were selected for Nanopore sequencing (Table 2).

Figure 3 shows percentages of reads mapping to viral reference, which were generally concordant with the

Figure 3

Comparison of Nanopore and Illumina results, as to proportions of reads mapping to the appropriate reference viral sequence, and proportions of reference genome sequenced at minimum 20-fold coverage, United Kingdom, 2017[a] (n = 8 samples)



Ct: cycle threshold; TC: template/complement.

[a] The Nanopore and Illumina data presented in the figure were obtained in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

The percentage of total reads mapping to the appropriate reference sequence is plotted in the upper panel. Lower panels display the percentage of the reference genome sequenced to a minimum depth of 20-fold in the data generated, in dark blue or dark green for the Illumina sequence data, in light blue or light green for Nanopore data (MinION_TC).

Table 3

Summary of Nanopore mapping data on chikungunya and dengue virus positive samples by real-time reverse transcription-PCR, United Kingdom, 2017[a] (n = 8 samples)

Sample	Ct value	Total reads	% reads mapping to appropriate viral sequence	% 20x coverage	20x genome length (nt)	% 10x coverage	Reference[b]	Reference size (nt)	Max de novo contig (nt)
CHIKV 1	14.7	267,171	65.1	98.57	11,658	99.2	CHIKV	11,826	5,263
CHIKV 3	21.4	1,891,028	85.1	99.76	11,798	99.9	CHIKV	11,826	10,793
CHIKV 4	22.7	216,493	11.6	94.11	11,130	97.2	CHIKV	11,826	4,256
CHIKV 9	30.08	3,481,358	4.08	100	11,826	100	CHIKV	11,826	9,860
DENV 1	16.3	284,622	71.3	99.9	10,719	99.9	DENV 1	10,735	8,281
DENV 2	16.9	203,700	72.1	99.6	10,692	99.6	DENV 1	10,735	10,157
DENV 6	21.2	1,377,721	71.1	99.9	10,634	99.9	DENV 4	10,649	7,877
DENV 11	28.7	1,111,566	2.9	95.3	10,226	96.3	DENV 1	10,735	4,699

CHIKV: chikungunya virus; Ct: cycle threshold; DENV: dengue virus.

[a] The Nanopore data presented in the table were obtained in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

[b] For DENV the serotype is also indicated.

Illumina data, although a slight decrease is observed across the range of Ct values. In the Nanopore data, the highest mapped read percentages observed were 85.12% and 72.14% for CHIKV3 and DENV 2 respectively, compared with 95.23% and 92.56% in the Illumina data from the same samples. While in high Ct samples the viral proportion drops to 4.08% for CHIKV 9 and 2.90% for DENV 11, from 6.66% and 3.73% in the Illumina data.

Despite the decrease in proportion of mapped viral reads, comparable genome coverage is observed at both 20x and 10x (Figure 3, Table 3) and is even increased compared with Illumina data at lower viral titres, e.g. 100% at 20x for CHIKV 9 compared with 95.98% in the Illumina data and 95.25% for the high Ct DENV 11 sample, which generated 94.44% coverage from the Illumina data. Average read lengths in Nanopore data ranged from 564 to 886 bp (Table 2).

Figure 4 shows coverage depth of reads mapped across the relevant genome for each sample sequenced by both Illumina and Nanopore. Read levels are not normalised thus actual depth is a function of total reads sequenced, but the pattern of coverage seen is highly similar suggesting it is more dependent upon the SISPA methodology than sequencing library preparation. From Nanopore consensus genome sequences, between 99.93% and 100% of bases called per sample agreed with the Illumina generated sequence.

Metagenomic data analysis and coinfection identification

To test the applicability of a metagenomic analysis approach to the data, we assessed read taxonomic classification using Kraken (Figure 5). The distribution of reads classified as CHIKV, DENV, other viruses, bacteria, and archaea/eukaryota show a similar pattern for Illumina and Nanopore data. The proportion of unclassified reads for each sample increased with Ct value, as the proportion of human origin reads is higher and the human genome is not represented in our Kraken database. A decrease in the percentage of CHIKV and DENV classified reads is observed for MinION data compared with Illumina, but was sufficient to identify the correct predominant virus in all samples.

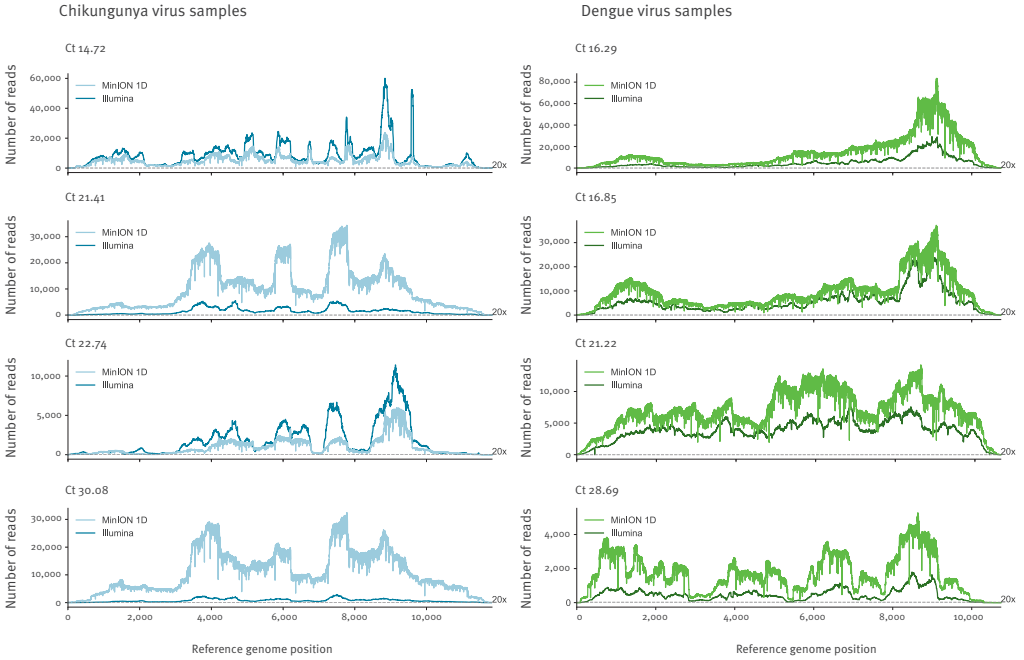
Kraken analysis also allowed for the identification of a DENV coinfection in sample CHIKV 3, the consensus sequence of which was unique in the sample set, eliminating cross-contamination from the DENV positive samples as potential source. Kraken classified 0.08% of Illumina reads and 0.15% of MinION reads as DENV. Using reference mapping to validate the finding, 0.22% of Illumina reads and 0.43% of MinION reads mapped to a DENV-1 reference genome. Genome coverage at 20x of 99.73% and 95.99% was achieved for the primary CHIKV and secondary DENV coinfection respectively, with a single MinION flow cell.

De novo assembly

De novo assembly of the data was attempted using Canu [42] and contigs identified using Basic Local Alignment Search Tool against a Nt database (BLASTn). Table 3 lists the longest viral contig length identified in each sample, ranging from 4.2 Kb (36% of reference genome size) to 10.8 Kb (91%) for CHIKV and 4.7 Kb (44%) to 10.1 Kb (95%) for DENV. Identification of the pathogen present without prior knowledge would have therefore been possible for all samples.

Figure 4

Coverage depth across the chikungunya or dengue viral genome, United Kingdom, 2017[a] (n = 8 samples)



Ct: cycle threshold.

[a] The Nanopore and Illumina data presented in the figure were obtained in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

Each graph corresponds to a given sample, defined by its Ct value. Read depth (y-axis) across the genome (x-axis) following reference alignment is shown. Illumina coverage is shown in darker blue and darker green for chikugunya and dengue virus positive samples respectively. Nanopore (MinION) coverage is indicated in lighter blue or lighter green for chikungunya and dengue virus positive samples respectively. Total depth has not been normalised; comparison is to show overall pattern of coverage is highly similar across the methods. Dotted horizontal line indicates depth of 20x coverage, used for consensus calling.

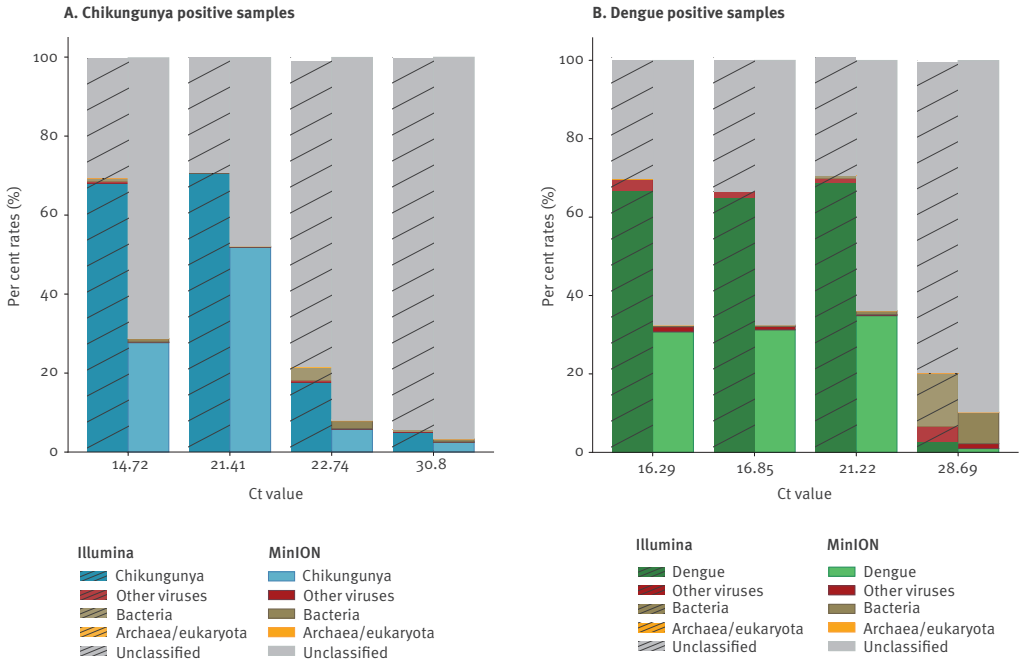
Updated MinION library kits

We repeated the sequencing of the coinfected CHIKV 3 sample using the MinION 1D² (SQK-LSK308) and Rapid (SQK-RBK001) kits, currently the most accurate and the fastest library preparation kits available, respectively. Using the 1D²kit 74.5% of reads generated mapped to CHIKV and 0.37% to DENV, while from the Rapid kit the result was 66.26% and 0.29% respectively (both lower than observed in the 2D chemistry). Coverage at 20x for CHIKV was above 99% for both kits, and for DENV was 95.04% from the 1D² and 81.09% from the Rapid kit (Table 4). Coverage depth pattern across the genome for both viruses (Figure 6) was similar for all library kits tested. Near-maximum coverage for both viruses was obtained within 30 min with the 2D kit, 8 min with the 1D² kit and 85 min with the Rapid kit (Supplementary Figure 1). De novo assembly (Table 4) produced best CHIKV contigs of 10.7, 11.3 and 11.4 Kb for the 2D, 1D² and Rapid libraries respectively and the longest contigs generated for DENV were 7.5, 2.2 and 4.2 Kb.

The 1D data from the Rapid kit was sufficient to call a consensus from 11,647/11,826 bases of the CHIKV reference with 179/11,826 bases called as ambiguous or too low coverage. All bases called were concordant with the Illumina consensus. A polishing step using Nanopolish [37] with a subset of the mapped reads (ca 100x coverage depth) significantly reduced ambiguous calls to 90/11,826, introducing a single disagreement with the Illumina consensus (99.99% concordance). Despite considerably greater read depth, the 1D² kit called only

Figure 5

Kraken classification of reads from metagenomic sequencing in (A) chikungunya and (B) dengue real-time reverse transcription-PCR positive samples, United Kingdom, 2017[a] (n = 8 samples)



Ct: cycle threshold.

[a] Read taxonomic classification using Kraken was conducted in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

Kraken classification distribution comparison for Illumina (cross-hatched) and Nanopore data. Reads grouped as either chikungunya virus (blue in panel A), dengue virus (green in panel B), other viruses (brown), archea/eukaryota (orange), bacteria (brown) or unclassified (grey).

11,082/11,826 due to a higher proportion, 744/11826, of ambiguous base calls, suggesting 1D reads are most suitable for this approach.

Discussion

These results clearly show that there are considerable levels of viral nucleic acid present in a large proportion of CHIKV and DENV qRT-PCR positive clinical samples, and demonstrate that relatively modest metagenomic sequencing is capable of elucidating significant portions of viral genome even for samples with Ct values at the higher end of clinical range. A decreased Ct value coincided with an increased proportion of viral reads, with a considerable level of variation between samples, likely because of the total level of non-viral host/background nucleic acid present due to variability between patients or in sample handling during collection, storage and testing. For example, the two lowest viral titre CHIKV samples (13 and 14) have similar Ct values (32.2 and 32.57) but varied significantly in the proportion of viral reads (5.03% and 21.72%). The 5.03% viral reads in CHIKV13 is the lowest for CHIKV, yet still sufficient to generate 88.5% of the CHIKV genome at 20x depth from just ca 662,000 paired-end Illumina reads. This amount of genomic information is highly informative and further sequencing would likely increase coverage. Only seven of the 73 total CHIKV diagnostic samples tested in 2016 had a Ct greater than 32.2 (including sample CHIKV14) (Table 1), which suggests that for the majority (> 90%) of CHIKV PCR positive samples, viral load is sufficient for genome sequencing directly from patient samples without further viral enrichment beyond a simple DNAse digestion (Figure 1). The lowest viral read proportion observed in the DENV samples was 0.47% in DENV12, Ct 31.29, which generated 71.5% coverage at 20x depth (increased to 77.8 at 10x

Table 4

Comparison of Nanopore mapping data across library kits, United Kingdom, 2017[a] (n = 8 samples)

Platform	Kit information	Flow cell (FLO-)	Virus identified	Total reads (nt)	% reads mapping	% 20x coverage	% 10x coverage	Reference[b]	Reference size (nt)	Max de novo contig (nt)
Illumina	Nextera XT	NA	CHIKV	1,391,258	95.23	98.86	99.37	CHIKV	11,826	7,321
Illumina	Nextera XT	NA	DENV	1,391,258	0.22	63.66	77.82	DENV1	10,735	6,613
MinION 2D	SQK-LSK208	MIN106	CHIKV	1,891,028	85.12	99.73	99.91	CHIKV	11,826	10,793
MinION 2D	SQK-LSK208	MIN106	DENV	1,891,028	0.43	95.99	96.09	DENV1	10,735	7,549
MinION 1D²	SQK-LSK308	MIN107	CHIKV	5,080,906	74.50	99.94	100	CHIKV	11,826	11,369
MinION 1D²	SQK-LSK308	MIN107	DENV	5,080,906	0.37	95.04	96.42	DENV1	10,735	2,199
MinION Rapid	SQK-RBK001	MIN106	CHIKV	611,110	66.26	99.66	99.68	CHIKV	11,826	11,473
MinION Rapid	SQK-RBK001	MIN106	DENV	611,110	0.29	81.09	90.83	DENV1	10,735	4,227

CHIKV: chikungunya virus; Ct: cycle threshold; DENV: dengue virus. NA: not applicable.

[a] Results presented in the table were obtained in 2017 on samples that had been collected and found positive for chikungunya or dengue virus by real-time reverse transcription-PCR in 2016.

[b] For DENV the serotype is also indicated.

depth) from just 687,000 paired end Illumina reads and allowed for DENV serotype identification. Only 62 of 368 DENV cases in 2016 had a higher Ct, predicting that > 80% of PCR positive DENV samples have a viral load sufficient for genome sequencing (Figure 1). These estimates are based on Ct range distribution from a single year, results may vary from year to year.

The high yield of viral sequences from clinical CHIKV and DENV samples make the exciting prospect of metagenomic MinION viral whole-genome-sequencing feasible, even for lower viral titre samples. Evaluating this on a representative subset of our samples demonstrates that viral read proportions are in general agreement with that seen for Illumina sequencing, predicting a similar proportion of qRT-PCR positive patient samples would be suitable for direct metagenomic sequencing on the MinION. Differences in precise proportions of viral reads seen between Illumina and MinION are likely due to inter-library variation. Differences in genome coverage achieved are due to both differences in total reads generated per sample (not normalised between platforms) as well as differences in average read length. Of the samples tested on the MinION, the lowest titre samples CHIKV 9 and DENV 11 both generated near complete genome coverage.

We repeated the sequencing of the coinfected CHIKV 3 sample using the MinION 1D² (SQK-LSK308) and Rapid (SQK-RBK001) kits. A reduction in viral proportion of total reads was observed compared with the 2D kit, which may be due partly to the extended storage time of the original samples before retesting. In the case of the 1D² kit, the lower proportion was outweighed by a substantial increase in total data generated per flow cell (5 M vs 1.8 M reads). For the Rapid kit, the total data produced should be considered in the light of the greatly simplified sample workflow and turnaround-time.

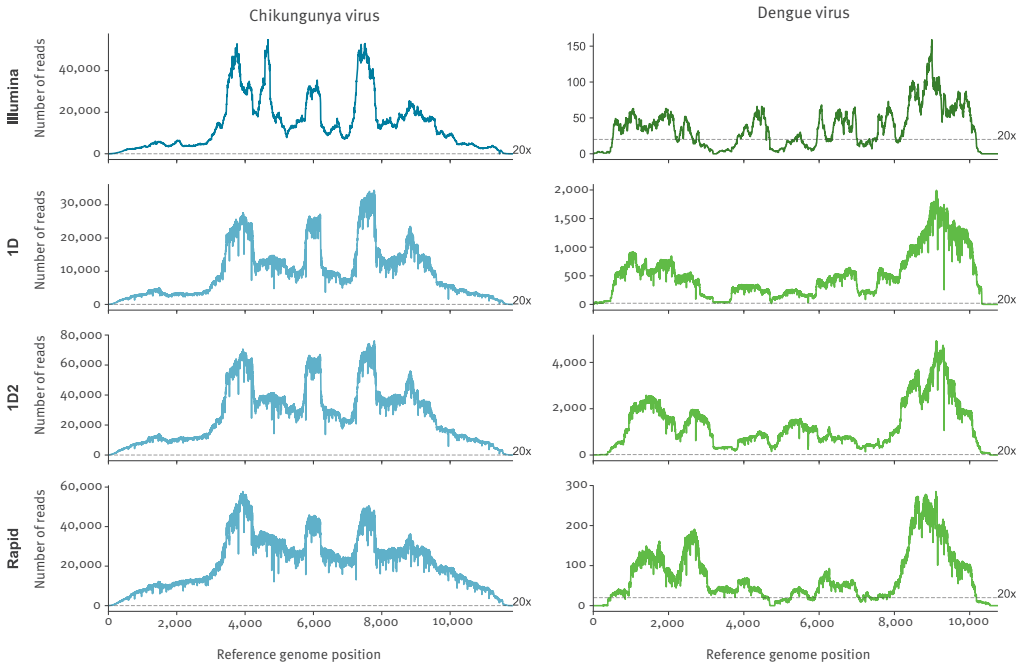
The use of metagenomics to elucidate genomic sequences of RNA viruses directly from clinical samples has several obvious benefits in public health applications. The primary benefit over targeted methods is the hypothesis-free nature of the assay, which allows identification and genomic characterisation of novel or unexpected RNA viral agents, either as primary or coinfectants (demonstrated here in the CHIKV/DENV coinfection sample), without any prior clinical knowledge. It also removes the need for laboratory optimisation of targeted methods, such as primer or bait-probe design and testing, and is not subject to escape mutations in target sites that afflict targeted sequencing and diagnostic methods. This issue particularly relevant for highly diverse RNA viruses, such as Lassa virus, which are difficult to assess using targeted methods, without regular reappraisal [43].

The principal limitation of the metagenomic approach is the limit of detection. The data generated here show that viral titres as low as 10^5 are sufficient for significant genome recovery by this method, but ZIKV is a recent example of a pathogen typically present at lower clinical titres, for which targeted methods are an absolute requirement [22,23]. For diagnostic purposes qRT-PCR has a lower limit of detection, provided the target site is conserved in the pathogen isolate tested. Clearly no single method is most suitable for both detection and genotyping of all pathogens and each has a role to play in differing circumstances.

The ability to generate genomic data directly from patient samples is clearly of great benefit to public

Figure 6

Comparison of genome coverage depth across the chikungunya virus or dengue virus genome for different sequencing library preparation methods in a sample coinfected with dengue and chikungunya viruses, United Kingdom, 2017[a] (n = 1 sample)



[a] Results presented in the Figure were obtained in 2017 on a sample that had been collected and found positive for chikungunya virus by real-time reverse transcription-PCR in 2016. In 2017, the sample was further found to be coinfected with CHIKV and DENV by metagenomic sequencing.

Read depth across both CHIKV and DENV genomes following reference alignment is shown for coinfection sample CHIKV 3, sequenced using four different sequencing library preparation/sequencing methods. Total coverage depth has not been normalised; comparison is to show overall pattern of coverage is highly similar across the methods. Dotted horizontal line indicates depth of 20x coverage, used for consensus calling.

health (reviewed in detail [44]). It can be used in a routine surveillance capacity or early during suspected outbreaks to link related cases who may be missed by traditional epidemiology [45] and identify outbreak cases distinct from typically circulating seasonal strains, which is key in regions endemic for the pathogen in question. The use of whole genome sequences offers the greatest precision for these applications, compared with typing methods based on specific genomic regions [44]. Whole genome sequencing on a portable device allows this information to be generated rapidly and within the affected region [24], enabling timely identification of an outbreak, or allaying fears of a potential one if cases are not linked. Furthermore mutations relating to viral drug resistance or pathogenicity can be monitored [44]. Therefore the ability to generate near-complete viral genome sequences directly from clinical samples on a portable sequencing device has many potential applications.

Conclusions

We demonstrate that across the clinically relevant range of viral loads an unexpectedly high proportion of reads generated metagenomically from CHIKV and DENV clinical samples are viral in origin. Therefore metagenomic sequencing provides an effective approach for the analysis of CHIKV and DENV genomes directly from the majority of qRT-PCR positive serum and plasma samples, without the need for culture or viral nucleic acid enrichment beyond a simple DNA digestion. We demonstrate this is equally possible on the Oxford Nanopore MinION, making metagenomic

whole genome sequencing potentially feasible in the field.

Acknowledgements

This work was funded via an NIHR HPRU in Emerging and Zoonotic Infections PhD studentship awarded to L. Kafetzopoulou. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health. Oxford Nanopore Technologies provided some reagents free of charge and funded author conference attendance.

Conflict of interest

Oxford Nanopore Technologies provided some reagents free of charge and funded author conference attendance.

Authors' contributions

Performed experiments: LEK, KL, AC, DC

Performed Data analysis: LEK, KE, STP

Design of study: LEK, EA, JO, RH, JAH, MWC, RV, STP

Wrote the manuscript: LEK, STP

All authors reviewed the manuscript.

References

1. Papa A. Emerging arboviral human diseases in Southern Europe. J Med Virol. 2017;89(8):1315-22. https://doi.org/10.1002/jmv.24803 PMID: 28252204
2. Gould E, Pettersson J, Higgs S, Charrel R, de Lamballerie X. Emerging arboviruses: Why today? One Health. 2017;4:1-13. https://doi.org/10.1016/j.onehlt.2017.06.001 PMID: 28785601
3. Weaver SC, Reisen WK. Present and future arboviral threats. Antiviral Res. 2010;85(2):328-45. https://doi.org/10.1016/j.antiviral.2009.10.008 PMID: 19857523
4. Thiberville S-D, Moyen N, Dupuis-Maguiraga L, Nougairede A, Gould EA, Roques P, et al. Chikungunya fever: epidemiology, clinical syndrome, pathogenesis and therapy. Antiviral Res. 2013;99(3):345-70. https://doi.org/10.1016/j.antiviral.2013.06.009 PMID: 23811281
5. World Health Organization (WHO). WHO publishes list of top emerging diseases likely to cause major epidemics. Geneva:WHO; 2017 Nov 10. [Accessed 26 Feb 2018]; Available from: http://www.who.int/medicines/ebola-treatment/WHO-list-of-top-emerging-diseases/en/
6. Burt FJ, Chen W, Miner JJ, Lenschow DJ, Merits A, Schnettler E, et al. Chikungunya virus: an update on the biology and pathogenesis of this emerging pathogen. Lancet Infect Dis. 2017;17(4):e107-17. https://doi.org/10.1016/S1473-3099(16)30385-1 PMID: 28159534
7. Mavalankar D, Shastri P, Bandyopadhyay T, Parmar J, Ramani KV. Increased mortality rate associated with chikungunya epidemic, Ahmedabad, India. Emerg Infect Dis. 2008;14(3):412-5. https://doi.org/10.3201/eid1403.070720 PMID: 18325255
8. Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1545-602. https://doi.org/10.1016/S0140-6736(16)31678-6 PMID: 27733282
9. Patterson J, Sammon M, Garg M. Dengue, Zika and Chikungunya: Emerging Arboviruses in the New World. West J Emerg Med. 2016;17(6):671-9. https://doi.org/10.5811/westjem.2016.9.30904 PMID: 27833670
10. Burt FJ, Rolph MS, Rulli NE, Mahalingam S, Heise MT. Chikungunya: a re-emerging virus. Lancet. 2012;379(9816):662-71. https://doi.org/10.1016/S0140-6736(11)60281-X PMID: 22100854
11. Simmons CP, Farrar JJ, van Vinh Chau N, Wills B. Dengue. N Engl J Med. 2012;366(15):1423-32. https://doi.org/10.1056/NEJMra1110265 PMID: 22494122
12. Furuya-Kanamori L, Liang S, Milinovich G, Soares Magalhaes RJ, Clements ACA, Hu W, et al. Co-distribution and co-infection of chikungunya and dengue viruses. BMC Infect Dis. 2016;16(1):84. https://doi.org/10.1186/s12879-016-1417-2 PMID: 26936191
13. Perera-Lecoin M, Luplertlop N, Surasombatpattana P, Liégeois F, Hamel R, Thongrungkiat S, et al. Dengue and Chikungunya Coinfection – The Emergence of an Underestimated Threat. In: Rodriguez-Morales AJ, editor. Current Topics in Chikungunya. InTech; 2016.
14. Omarjee R, Prat C, Flusin O, Boucau S, Tenebray B, Merle O, et al. Importance of case definition to monitor ongoing outbreak of chikungunya virus on a background of actively circulating dengue virus, St Martin, December 2013 to January 2014. Euro Surveill. 2014;19(13):20753. https://doi.org/10.2807/1560-7917.ES2014.19.13.20753 PMID: 24721537
15. Brito CAA, Azevedo F, Cordeiro MT, Marques ETA Jr, Franca RFO. Central and peripheral nervous system involvement caused by Zika and chikungunya coinfection. PLoS Negl Trop Dis. 2017;11(7):e0005583. https://doi.org/10.1371/journal.pntd.0005583 PMID: 28704365
16. Wilder-Smith A, Gubler DJ, Weaver SC, Monath TP, Heymann DL, Scott TW. Epidemic arboviral diseases: priorities for research and public health. Lancet Infect Dis. 2017;17(3):e101-6. https://doi.org/10.1016/S1473-3099(16)30518-7 PMID: 28011234
17. Briese T, Paweska JT, McMullan LK, Hutchison SK, Street C, Palacios G, et al. Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. PLoS Pathog. 2009;5(5):e1000455. https://doi.org/10.1371/journal.ppat.1000455 PMID: 19478873
18. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, Reeder SA, et al. Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. PLoS Pathog. 2008;4(11):e1000212. https://doi.org/10.1371/journal.ppat.1000212 PMID: 19023410
19. Grard G, Fair JN, Lee D, Slikas E, Steffen I, Muyembe J-J, et al. A novel rhabdovirus associated with acute hemorrhagic fever in central Africa. PLoS Pathog. 2012;8(9):e1002924. https://doi.org/10.1371/journal.ppat.1002924 PMID: 23028323
20. Metsky HC, Matranga CB, Wohl S, Schaffner SF, Freije CA, Winnicki SM, et al. Zika virus evolution and spread in the Americas. Nature. 2017;546(7658):411-5. https://doi.org/10.1038/nature22402 PMID: 28538734
21. Grubaugh ND, Ladner JT, Kraemer MUG, Dudas G, Tan AL, Gangavarapu K, et al. Genomic epidemiology reveals multiple introductions of Zika virus into the United States. Nature. 2017;546(7658):401-5. https://doi.org/10.1038/nature22400 PMID: 28538723
22. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546(7658):406-10. https://doi.org/10.1038/nature22401 PMID: 28538727
23. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. Nat Protoc. 2017;12(6):1261-76. https://doi.org/10.1038/nprot.2017.066 PMID: 28538739
24. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530(7589):228-32. https://doi.org/10.1038/nature16996 PMID: 26840485
25. Walter MC, Zwirglmaier K, Vette P, Holowachuk SA, Stoecker K, Genzel GH, et al. MinION as part of a biomedical rapidly deployable laboratory. J Biotechnol. 2017;250:16-22. https://doi.org/10.1016/j.jbiotec.2016.12.006 PMID: 27939320
26. Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB. MinION™ nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience. 2017;6(3):1-10. https://doi.org/10.1093/gigascience/gix007 PMID: 28327976
27. Batovska J, Lynch SE, Rodoni BC, Sawbridge TI, Cogan NO. Metagenomic arbovirus detection using MinION nanopore sequencing. J Virol Methods. 2017;249:79-84. https://doi.org/10.1016/j.jviromet.2017.08.019 PMID: 28855093
28. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med. 2015;7(1):99. https://doi.org/10.1186/s13073-015-0220-9 PMID: 26416663

29. Sardi SI, Somasekar S, Naccache SN, Bandeira AC, Tauro LB, Campos GS, et al. Coinfections of Zika and Chikungunya Viruses in Bahia, Brazil, Identified by Metagenomic Next-Generation Sequencing. J Clin Microbiol. 2016;54(9):2348-53. https://doi.org/10.1128/JCM.00877-16 PMID: 27413190

30. Drosten C, Göttig S, Schilling S, Asper M, Panning M, Schmitz H, et al. Rapid detection and quantification of RNA of Ebola and Marburg viruses, Lassa virus, Crimean-Congo hemorrhagic fever virus, Rift Valley fever virus, dengue virus, and yellow fever virus by real-time reverse transcription-PCR. J Clin Microbiol. 2002;40(7):2323-30. https://doi.org/10.1128/JCM.40.7.2323-2330.2002 PMID: 12089242

31. Edwards CJ, Welch SR, Chamberlain J, Hewson R, Tolley H, Cane PA, et al. Molecular diagnosis and analysis of Chikungunya virus. J Clin Virol. 2007;39(4):271-5. https://doi.org/10.1016/j.jcv.2007.05.008 PMID: 17627877

32. Loman NJ, Quinlan AR. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics. 2014;30(23):3399-401. https://doi.org/10.1093/bioinformatics/btu555 PMID: 25143291

33. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

34. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

35. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

36. Penedos AR, Myers R, Hadef B, Aladin F, Brown KE. Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks. PLoS One. 2015;10(11):e0143081. https://doi.org/10.1371/journal.pone.0143081 PMID: 26569100

37. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods. 2015;12(8):733-5. https://doi.org/10.1038/nmeth.3444 PMID: 26076426

38. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46. https://doi.org/10.1186/gb-2014-15-3-r46 PMID: 24580807

39. Lewandowski K, Bell A, Miles R, Carne S, Wooldridge D, Manso C, et al. The Effect of Nucleic Acid Extraction Platforms and Sample Storage on the Integrity of Viral RNA for Use in Whole Genome Sequencing. J Mol Diagn. 2017;19(2):303-12. https://doi.org/10.1016/j.jmoldx.2016.10.005 PMID: 28041870

40. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455-77. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599

41. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27(4):578-9. https://doi.org/10.1093/bioinformatics/btq683 PMID: 21149342

42. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722-36. https://doi.org/10.1101/gr.215087.116 PMID: 28298431

43. Andersen KG, Shapiro BJ, Matranga CB, Sealfon R, Lin AE, Moses LM, et al. Viral Hemorrhagic Fever Consortium. Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell. 2015;162(4):738-50. https://doi.org/10.1016/j.cell.2015.07.020 PMID: 26276630

44. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. Nat Rev Microbiol. 2017;15(3):183-92. https://doi.org/10.1038/nrmicro.2016.182 PMID: 28090077

45. Keita M, Duraffour S, Loman NJ, Rambaut A, Diallo B, Magassouba N, et al. Unusual Ebola Virus Chain of Transmission, Conakry, Guinea, 2014-2015. Emerg Infect Dis. 2016;22(12):2149-52. https://doi.org/10.3201/eid2212.160847 PMID: 27869596

VIROLOGY

Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak

L. E. Kafetzopoulou[1,2,3], S. T. Pullan[1,2], P. Lemey[4], M. A. Suchard[5], D. U. Ehichioya[3,6], M. Pahlmann[3,6], A. Thielebein[3,6], J. Hinzmann[3,6], L. Oestereich[3,6], D. M. Wozniak[3,6], K. Efthymiadis[7], D. Schachten[3], F. Koenig[3], J. Matjeschk[3], S. Lorenzen[3], S. Lumley[1], Y. Ighodalo[8], D. I. Adomeh[8], T. Olokor[8], E. Omomoh[8], R. Omiunu[8], J. Agbukor[8], B. Ebo[8], J. Aiyepada[8], P. Ebhodaghe[8], B. Osiemi[8], S. Ehikhametalor[8], P. Akhilomen[8], M. Airende[8], R. Esumeh[8], E. Muoebonam[8], R. Giwa[8], A. Ekanem[8], G. Igenegbale[8], G. Odigie[8], G. Okonofua[8], R. Enigbe[8], J. Oyakhilome[8], E. O. Yerumoh[8], I. Odia[8], C. Aire[8], M. Okonofua[8], R. Atafo[8], E. Tobin[8], D. Asogun[8,9], N. Akpede[8], P. O. Okokhere[8,9], M. O. Rafiu[8], K. O. Iraoyah[8], C. O. Iruolagbe[8], P. Akhideno[8], C. Erameh[8], G. Akpede[8,9], E. Isibor[8], D. Naidoo[10], R. Hewson[1,2,11,12], J. A. Hiscox[2,13,14], R. Vipond[1,2], M. W. Carroll[1,2], C. Ihekweazu[15], P. Formenty[10], S. Okogbenin[8,9], E. Ogbaini-Emovon[8]*, S. Günther[3,6]*†, S. Duraffour[3,6]*

The 2018 Nigerian Lassa fever season saw the largest ever recorded upsurge of cases, raising concerns over the emergence of a strain with increased transmission rate. To understand the molecular epidemiology of this upsurge, we performed, for the first time at the epicenter of an unfolding outbreak, metagenomic nanopore sequencing directly from patient samples, an approach dictated by the highly variable genome of the target pathogen. Genomic data and phylogenetic reconstructions were communicated immediately to Nigerian authorities and the World Health Organization to inform the public health response. Real-time analysis of 36 genomes and subsequent confirmation using all 120 samples sequenced in the country of origin revealed extensive diversity and phylogenetic intermingling with strains from previous years, suggesting independent zoonotic transmission events and thus allaying concerns of an emergent strain or extensive human-to-human transmission.

Lassa fever is an acute viral hemorrhagic illness, first described in 1969 in the town of Lassa, Nigeria (*1*). It is contracted primarily through exposure to urine or feces of infected *Mastomys* spp. rodents or, less frequently, through the bodily fluids of infected humans. Lassa virus (LASV) is endemic in parts of West Africa, including Nigeria, Benin, Côte d'Ivoire, Mali, Sierra Leone, Guinea, and Liberia (*2*). The upsurge of Lassa fever cases during the 2018 endemic season in Nigeria—referred to here as the 2018 Lassa fever outbreak—has been the largest on record, reaching 1495 suspected cases and 376 confirmed cases and affecting more than 18 states by 18 March (fig. S1). This notably exceeds the 102 confirmed cases reported during the same period in 2017 (fig. S1) (*3*). The unprecedented scale of the outbreak raised fears of the emergence of a strain with a higher rate of transmission. Because of these concerns, on 28 February the Nigeria Centre for Disease Control (NCDC) and the World Health Organization (WHO) urgently requested sequencing information and preliminary results from our pilot-scale study, in which we used a metagenomic approach with the Oxford Nanopore MinION device (Oxford Nanopore Technologies) to conduct in-country, mid-outbreak viral genome sequencing. This instigated a major uptick in sequencing efforts, leading to the sequencing of 120 samples.

Nanopore sequencing is an emerging technology with great potential. The MinION is a small, robust sequencing device suited for the genetic analysis of pathogens in remote or resource-limited settings (*4*). Nanopore sequencing of polymerase chain reaction (PCR) amplicons of Ebola virus genomes provided important data from the field in real time during the 2014–2016 Ebola virus disease outbreak in West Africa (*5*), and a more sophisticated multiplex amplicon sequencing methodology (*6*) has been used effectively during recent Zika and yellow fever outbreaks in Brazil (*7*, *8*). However, highly variable pathogens such as LASV present a substantial challenge for this type of amplicon-based approach. Owing to an interstrain nucleic acid sequence variation of up to 32 and 25% for the L (large segment encoding the RNA polymerase and the zinc-binding protein) and S (small segment encoding the glycoprotein and the nucleoprotein) segments, respectively (*9*), even PCR-based laboratory diagnosis poses a serious challenge. Designing targeted whole-genome sequencing approaches, such as those using PCR amplicons or bait-and-capture probes, without prior knowledge of the targeted LASV lineage is therefore cumbersome. Random reverse-transcription (RT) and amplification by sequence-independent single primer amplification (SISPA) for metagenomic sequencing to identify RNA viruses has been demonstrated to work on the MinION (*10*), and our previous work highlighted the feasibility of retrieving complete viral genomes directly from patient samples at clinically relevant viral titers using this approach for dengue and chikungunya viruses (*11*). We describe here the application of field metagenomic sequencing of LASV at the Irrua Specialist Teaching Hospital (ISTH), Edo State, during the 2018 Lassa fever season.

A total of 120 LASV-positive samples were sequenced during a 7-week mission; these were selected on the basis of cycle threshold value and location of the 341 cases reported by ISTH between 1 January and 18 March 2018 (figs. S1 and S2). The majority of samples originated from Edo State followed by Ondo and Ebonyi (fig. S2). Selected samples covered the wide range of clinical viral loads observed, including several samples testing negative in one of the two real-time RT-PCR assays used (fig. S3 and data S1). Up to six samples were run in multiplex per MinION flow cell, along with a negative control. To produce high-confidence consensus sequences for phylogenetic inference, we chose to map both basecalled reads and raw signal data to a reference sequence and call variants using Nanopolish software, as developed for the West African Ebola virus disease outbreak (*5*); basecalled reads were then remapped to the consensus and a further round of correction was applied (fig. S4). Owing to the diversity of LASV, selection of an individual reference genome for read alignment was required for each sample. To select the closest existing LASV reference genome, nonhuman reads from each sample were assembled de novo using Canu (*12*). A notable proportion of reads generated per sample were LASV at an

[1]Public Health England, National Infection Service, Porton Down, UK. [2]National Institute of Health Research (NIHR), Health Protection Research Unit in Emerging and Zoonotic Infections, University of Liverpool, Liverpool, UK. [3]Bernhard Nocht Institute for Tropical Medicine, Hamburg, Germany. [4]Department of Microbiology and Immunology, Rega Institute, KU Leuven – University of Leuven, Leuven, Belgium. [5]Departments of Biomathematics, Biostatistics, and Human Genetics, University of California, Los Angeles, CA, USA. [6]German Center for Infection Research (DZIF), partner site Hamburg, Germany. [7]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium. [8]Irrua Specialist Teaching Hospital, Irrua, Nigeria. [9]Faculty of Clinical Sciences, College of Medicine, Ambrose Alli University, Ekpoma, Nigeria. [10]World Health Organization, Geneva, Switzerland. [11]Faculty of Infectious and Tropical Diseases, Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, London, UK. [12]Faculty of Clinical Sciences and International Public Health, Liverpool School of Tropical Medicine, Liverpool, UK. [13]Singapore Immunology Network, Agency for Science, Technology and Research (A*STAR), Singapore. [14]Institute of Infection and Global Health, University of Liverpool, Liverpool, UK. [15]Nigeria Centre for Disease Control, Abuja, Nigeria.
*These authors contributed equally to this work.
†Corresponding author. Email: guenther@bni.uni-hamburg.de

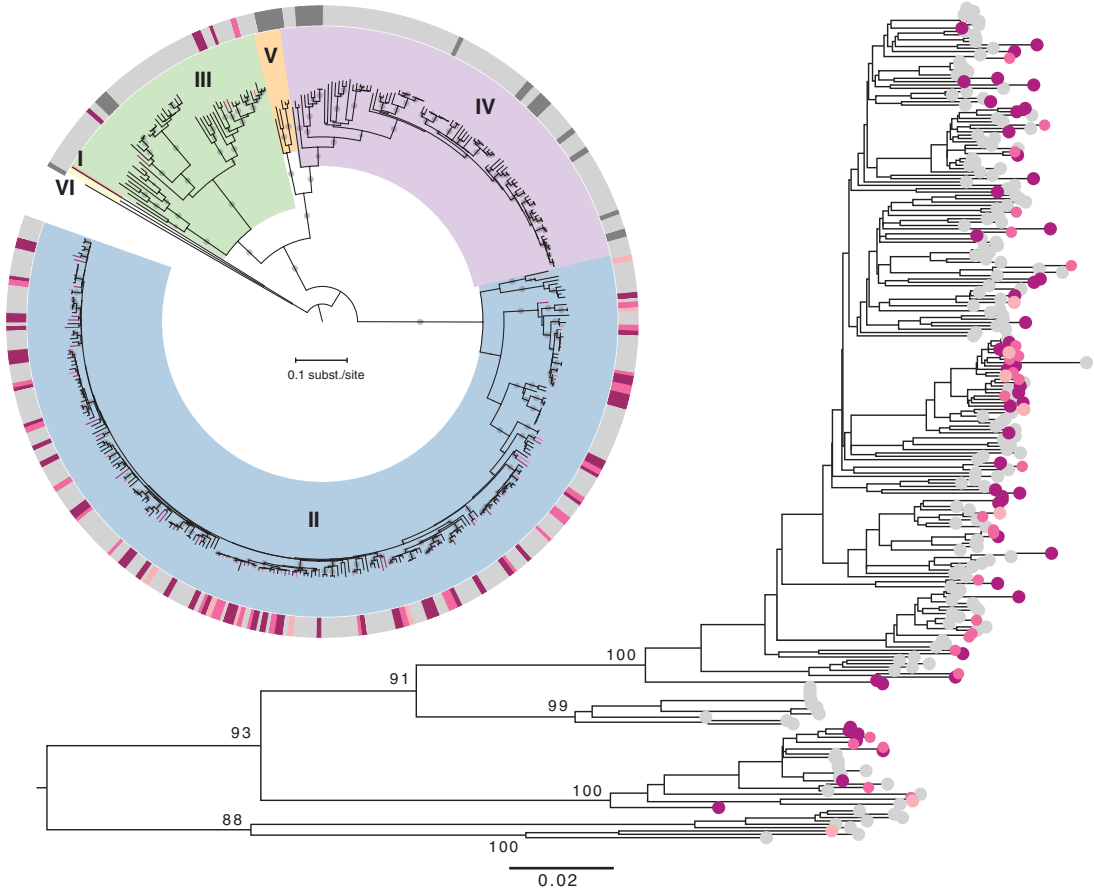


Fig. 1. Phylogenetic reconstruction of the S segment data. The circular tree includes 96 sequences from 2012 to 2017, 88 sequences from 2018, and sequences available from GenBank. The rectangular tree focuses on the genotype II clade (in blue in the circular tree), which includes most of the 2018 sequences. The six genotypes are indicated with different colors and roman numerals. Bootstrap support >90% is indicated with a small gray circle at the middle of their respective branches. The color strip highlights the human LASV sequences obtained from previous years (light gray); sequences obtained from rodent samples (dark gray); and, for 2018, the first seven sequences generated in Nigeria (light pink), the remaining 28 sequences analyzed on-site (medium pink), and the remaining sequences finalized in Europe (dark pink). The same color code is used in the genotype II rectangular tree. Bootstrap values >80% are shown for the major genotype II lineages.

average frequency of 4.26% with a maximum of 42.9%, allowing for sufficient genomic sequence (>70%) for phylogenetic comparison of at least one segment in 91 of the samples tested (figs. S3 to S6).

Additionally, sequences were validated by Illumina resequencing of 14 SISPA preparations, which matched with their Oxford Nanopore counterparts with little to no divergence, confirming the accuracy of the Oxford Nanopore approach (table S1).

Metagenomic classification using the Centrifuge software system (*13*) identified 0.10% of reads from sample 110 as originating from hepatitis A virus, providing 74% genome coverage at 20-fold depth. LASV accounted for 0.83% of reads in the same sample, providing 96% genome coverage. These findings demonstrate the potential of this simple approach to identify multiple RNA viruses, including those present as co-infections. In all other samples tested, LASV was the sole pathogen identified despite a small number of reads classified as other viruses (fig. S7 and data S1).

To dissect the molecular epidemiology of the 2018 Lassa fever outbreak in Nigeria, we performed phylogenetic analysis of all newly generated LASV sequences together with unpublished sequences from previous years (data S2) and sequences available in GenBank. We used this as a frame of reference to document how the genomic data generated in real time (made publicly available at virological.org) provided valuable epidemiological insights into the unfolding outbreak dynamics.

Maximum likelihood phylogenetic reconstruction of the S segment sequences indicates that all 2018 viruses fall within the Nigerian LASV diversity, specifically within genotypes II and III, and they are phylogenetically interspersed with Nigerian LASV sequences from previous years (Fig. 1). This phylogenetic pattern is mimicked by the L segment reconstruction (fig. S8). Only seven viruses in the entire genome dataset (n = 348) were identified as clustering

significantly differently in the L and S segments (supplementary methods), which is in line with the small number of potential LASV reassortments identified previously (*9*). The phylogenetic pattern implicates independent spillover from rodent hosts as the major driver of Lassa fever incidence during the outbreak (Fig. 1 and fig. S8).

However, a number of sequences from the 2018 outbreak clustered as pairs in the phylogenetic reconstructions, raising concerns over human-to-human transmission. We illustrate such cluster pairs in a Bayesian time-measured tree estimated from genotype II S (Fig. 2) and L segment sequences (fig. S9). These analyses resulted in highly similar evolutionary rate estimates for both segments (mean, ~1.2×10^{-3} substitutions per site per year) (Fig. 2 and figs. S9 and S10), in agreement with previous estimates (*9*). We used these rate estimates together with an estimate of the time between successive cases in a transmission chain to assess how many substitutions can be expected between directly linked infections. We compared conservative to more liberal expectations, the latter accommodating an independent upper estimate of potential sequencing errors (Fig. 2 and fig. S9). In the S segment, for example, more than two substitutions between sequences from directly linked infections is highly unlikely ($P < 0.01$ and $P = 0.03$, respectively, for the conservative and liberal probability estimates). This expectation is consistent with the low number of substitutions observed in the coding region of human-to-human LASV transmission (*14*). Four clusters of sequences showing ≤4 and ≤12 nucleotide differences in the S and L segments, respectively, were identified (035-045, 035-058, 137-138, and 053-089-106; for some of them, only the S or L segment sequence was available). Retrospective tracing revealed that the sequences for pairs 137-138 and 035-058 were derived from the same patients. Epidemiological investigation of the remaining clusters did not provide evidence for transmission chains, though direct linkage cannot be excluded. Even when applying liberal assumptions for the number of mutations during human-to-human transmission, the vast majority of cases during the 2018 outbreak resulted from spillover from the natural reservoir.

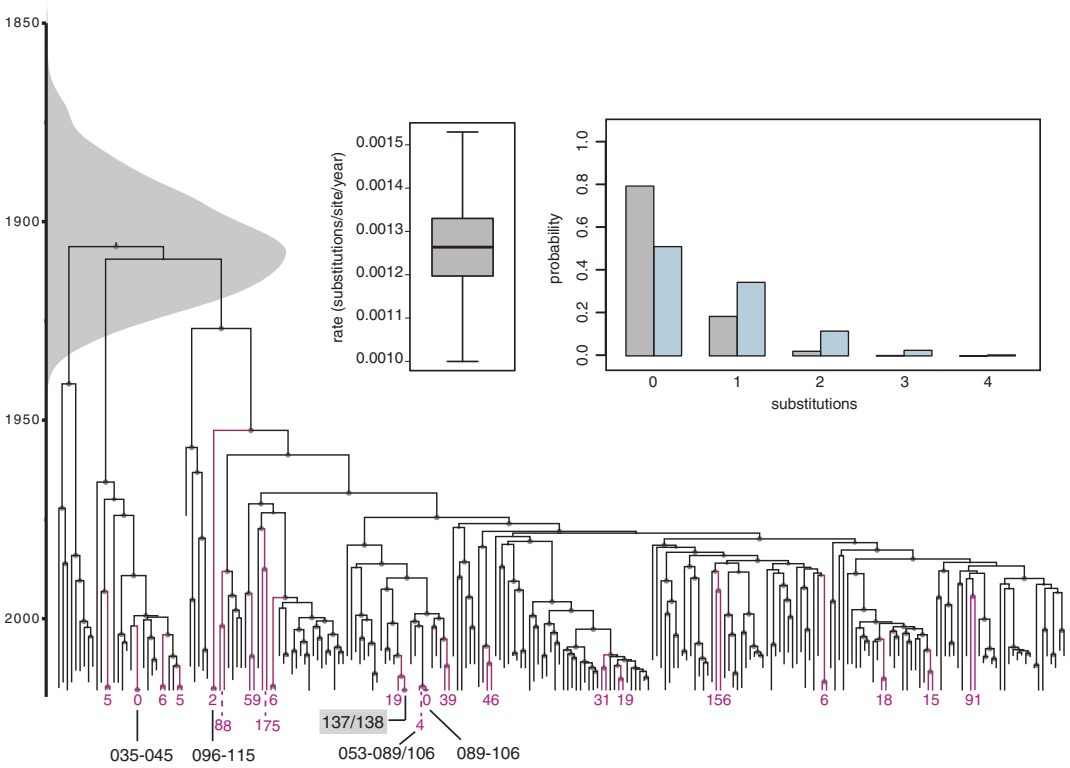


Fig. 2. Assessing the potential for direct linkage between pairs of 2018 sequences in the S segment. The maximum clade credibility tree summarizes a Bayesian evolutionary inference for the genotype II sequences in the S segment. A time scale and a marginal posterior distribution for the time to the most recent common ancestor are shown to the left. The size of the internal node circles reflects posterior probability support values. 2018 sequences clustering as pairs are indicated in dark pink; the number of substitutions between them is indicated at their respective tips. A posterior estimate of the evolutionary rate and probability distributions for observing a given number of substitutions during a human-to-human transmission event are shown as insets. The distribution represented by gray bars is based on the mean evolutionary rate estimate and a mean estimate for the generation time, whereas the light blue distribution is based on upper estimates and also incorporates an upper estimate for the MinION sequencing error (supplementary methods). At the bottom of the tree, clusters of sequences for which human-to-human transmission cannot be excluded according to the upper estimates of generation time are indicated. A pair of identical sequences (137-138) that was retrospectively found to be derived from the same patient is marked with a gray box. One pair (096-115) was disregarded as a potential transmission chain because of 21 differences in the L segment (fig. S9). The temporal signal before BEAST inference was explored in fig. S10.

A request for information on circulating strains was made on 28 February at the height of the outbreak; within 10 days, our pilot study was expedited and the initial analysis completed. The fact that the 2018 outbreak was fueled by the circulating LASV diversity and not by transmission of a new or divergent lineage was already evident from the first seven genomes generated by 10 March (fig. S1). This information was promptly communicated to the NCDC, forming the basis of its report released on 12 March 2018 (*15*). Whereas this small sample was restricted to genotype II, the final collection of 36 LASV genome sequences generated on-site also included a representative of genotype III (Fig. 1 and fig. S9), further supporting the spillover of long-standing LASV diversity in the outbreak. The conclusions drawn from the first set of genome sequences immediately eased fears of extensive human-to-human transmission and allowed public health resources to be allocated appropriately. The response was focused on intensified community engagement on rodent control, environmental sanitation, and safe food storage. Further research is needed to evaluate whether improved diagnostics and disease awareness and/or ecological and climate factors promoting transmission are the drivers behind the changing epidemiology of Lassa fever in Nigeria.

Portable metagenomic sequencing of genetically diverse RNA viruses on the MinION, direct from patient samples without the need to export material outside of the country of origin and with no pathogen-specific enrichment, is shown to be a feasible methodology enabling a real-time characterization of potential outbreaks in the field.

REFERENCES AND NOTES

1. J. D. Frame, J. M. Baldwin Jr., D. J. Gocke, J. M. Troup, *Am. J. Trop. Med. Hyg.* **19**, 670–676 (1970).
2. D. A. Asogun *et al.*, *PLOS Negl. Trop. Dis.* **6**, e1839 (2012).
3. WHO, "Lassa Fever – Nigeria" (2018); www.who.int/csr/don/23-march-2018-lassa-fever-nigeria/en/.
4. M. Jain, H. E. Olsen, B. Paten, M. Akeson, *Genome Biol.* **17**, 239 (2016).
5. J. Quick *et al.*, *Nature* **530**, 228–232 (2016).
6. J. Quick *et al.*, *Nat. Protoc.* **12**, 1261–1276 (2017).
7. N. R. Faria *et al.*, *Nature* **546**, 406–410 (2017).
8. N. R. Faria *et al.*, *Science* **361**, 894–899 (2018).
9. K. G. Andersen *et al.*, *Cell* **162**, 738–750 (2015).
10. A. L. Greninger *et al.*, *Genome Med.* **7**, 99 (2015).
11. L. E. Kafetzopoulou *et al.*, *Euro Surveill.* **23**, 1800228 (2018).
12. S. Koren *et al.*, *Genome Res.* **27**, 722–736 (2017).
13. D. Kim, L. Song, F. P. Breitwieser, S. L. Salzberg, *Genome Res.* **26**, 1721–1729 (2016).
14. S. L. M. Whitmer *et al.*, *Emerg. Infect. Dis.* **24**, 599–602 (2018).
15. Nigeria Centre for Disease Control, "Early Results of Lassa Virus Sequencing & Implications for Current Outbreak Response in Nigeria" (2018); https://ncdc.gov.ng/news/121/early-results-of-lassa-virus-sequencing-%26-implications-for-current-outbreak-response-in-nigeria.
16. P. Lemey, ISTH-BNITM-PHE/LASVsequencing: LASVrelease, Zenodo (2018); http://doi.org/10.5281/zenodo.1481015.

ACKNOWLEDGMENTS

We thank the health authorities of Nigeria for their cooperation during the outbreak response. **Funding:** L.E.K., S.T.P., R.H., R.V., M.W.C., and J.A.H. acknowledge funding by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emerging and Zoonotic Infections at the University of Liverpool in partnership with Public Health England (PHE), in collaboration with Liverpool School of Tropical Medicine. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health, or Public Health England. L.E.K. has received travel expenses and accommodation from Oxford Nanopore to speak at conferences regarding this work. L.E.K. has received some reagents free of charge from Oxford Nanopore in support of her Ph.D. project. M.W.C. has received reagents free of charge from Oxford Nanopore in support of previous projects not related to the work presented in this manuscript. L.E.K. and M.W.C. have not received other financial compensation nor hold shares. P.L. and M.A.S. acknowledge funding from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant 725422-ReservoirDOCS) and from the Wellcome Trust Collaborative Award, 206298/Z/17/Z. P.L. acknowledges support by the Special Research Fund, KU Leuven ("Bijzonder Onderzoeksfonds," KU Leuven, OT/14/115), and the Research Foundation–Flanders ("Fonds voor Wetenschappelijk Onderzoek – Vlaanderen," G066215N, G0D5117N, and G0B9317N). M.A.S. acknowledges support under National Science Foundation grant DMS 1264153. This study was supported by the German Federal Ministry of Health through support of the WHO Collaborating Centre for Arboviruses and Hemorrhagic Fever Viruses at the Bernhard Nocht Institute for Tropical Medicine (agreements ZMV I 1-2517WHO005 and ZMV I 1-2517WHO010) and through the Global Health Protection Program (agreement ZMVI1-2517-GHP-704), the German Federal Ministry for Economic Cooperation and Development through the Rapid Deployment Expert Group to Combat Threats (SEEG), the European Union's Horizon 2020 research and innovation program to S.G. (grant 653316-EVAg), and the German Research Foundation (DFG) to S.G. and D.U.E. (GU 883/4-1). D.U.E. acknowledges fellowships from Alexander von Humboldt Foundation and Kirmser Foundation. The funders had no role in the design and interpretation of the data and preparation of the manuscript. **Author contributions:** L.E.K., S.G., S.D., S.T.P., and P.L. conceptualized the study; L.E.K., S.T.P., and P.L. set up the methodology; L.E.K., J.H., A.T., S.D., and D.U.E. performed sequencing and data validation; L.E.K., P.L., M.A.S., S.T.P., D.S., F.K., J.M., and S.Lo. performed the formal sequencing data analysis; L.E.K., S.D., J.H., A.T., M.P., and L.O. performed sample selection, data collection, and organization of sequencing datasets; D.M.W., K.E., D.S., F.K., and J.M. set up and assisted with the bioinformatics pipeline; M.A.S., D.U.O., M.P., L.O., Y.I., D.I.A., T.O., E.O., R.O., J.Ag., B.E., J.Ai., P.E., B.O., S.E., P.A., M.A., R.Es., E.M., R.G., A.E., G.I., G.Od., G.Ok., R.En., J.O., E.O.Y., I.O., C.A., M.O., R.A., E.T., D.A., N.A., P.O.O., M.O.R., K.O.I., C.O.I., P.A., C.E., G.A., and E.I. performed diagnostic analysis; L.E.K., S.T.P., P.L., and S.D. visualized data presentation; L.E.K., S.T.P., P.L., and S.D. wrote the manuscript; all authors reviewed and edited the manuscript; S.G., M.W.C., J.A.H., R.H., and R.V. supervised the study; M.P., R.V., A.T., C.I., P.F., D.N., S.O., E.O.E., S.G., S.D., and S.Lu. performed project administration and implementation; S.G., P.L., M.W.C., R.V., R.H., J.A.H., L.E.K., and D.U.E. were involved in funding acquisition. **Competing interests:** C.I. is a member of the WHO Strategic Technical Advisory Group on Infectious Diseases; D.A. serves as an expert for the WHO R&D Blueprint for action to prevent epidemics (the Blueprint); S.G. is a member of the Scientific Advisory Group (SAG) to advise WHO on the implementation of the Blueprint, including a plan for international coordination of the R&D effort in the event of a highly infectious pathogen epidemic; S.O. serves as an expert for the Blueprint. All other authors declare no competing interests. **Data and materials availability:** LASV sequences from 2018 are deposited in GenBank under BioProject PRJNA482058 (data S1); sequences from 2012 to 2017 are deposited under BioProjects PRJNA482054 and PRJNA482058 (data S2). Alignments, trees, and BEAST xml files are available at https://github.com/ISTH-BNITM-PHE/LASVsequencing and in (*16*).

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/363/6422/74/suppl/DC1
Materials and Methods
Figs. S1 to S10
Table S1
References (*17*–*30*)
Data S1 and S2

3 August 2018; accepted 12 November 2018
10.1126/science.aau9343

Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak

L. E. Kafetzopoulou, S. T. Pullan, P. Lemey, M. A. Suchard, D. U. Ehichioya, M. Pahlmann, A. Thielebein, J. Hinzmann, L. Oestereich, D. M. Wozniak, K. Efthymiadis, D. Schachten, F. Koenig, J. Matjeschk, S. Lorenzen, S. Lumley, Y. Ighodalo, D. I. Adomeh, T. Olokor, E. Omomoh, R. Omiunu, J. Agbukor, B. Ebo, J. Aiyepada, P. Ebhodaghe, B. Osiemi, S. Ehikhametalor, P. Akhilomen, M. Airende, R. Esumeh, E. Muoebonam, R. Giwa, A. Ekanem, G. Igenegbale, G. Odigie, G. Okonofua, R. Enigbe, J. Oyakhilome, E. O. Yerumoh, I. Odia, C. Aire, M. Okonofua, R. Atafo, E. Tobin, D. Asogun, N. Akpede, P. O. Okokhere, M. O. Rafiu, K. O. Iraoyah, C. O. Iruolagbe, P. Akhideno, C. Erameh, G. Akpede, E. Isibor, D. Naidoo, R. Hewson, J. A. Hiscox, R. Vipond, M. W. Carroll, C. Ihekweazu, P. Formenty, S. Okogbenin, E. Ogbaini-Emovon, S. Günther and S. Duraffour

Science **363** (6422), 74-77.
DOI: 10.1126/science.aau9343

Mobile detection of Lassa virus

Lassa fever is a hemorrhagic viral disease endemic to West Africa. Usually, each year sees only a smattering of cases reported, but hospitalized patients risk a 15% chance of death. Responding to fears that a 10-fold surge in cases in Nigeria in 2018 signaled an incipient outbreak, Kafetzopoulou *et al.* performed metagenomic nanopore sequencing directly from samples from 120 patients (see the Perspective by Bhadelia). Results showed no strong evidence of a new strain emerging nor of person-to-person transmission; rather, rodent contamination was the main source. To prevent future escalation of this disease, we need to understand what triggers the irruption of rodents into human dwellings.

Science, this issue p. 74; see also p. 30